Université du Québec
Institut national de la recherche scientifique
Centre Enérgie Matériaux et Télécommunications

ENERGY EFFICIENCY AND QOS MAXIMIZATION IN FUTURE GREEN WIRELESS HETNETS WITH OPTIMIZED BASE-STATION POWER SUPPLY ON/OFF SWITCHING STRATEGIES

Par

Mariam Lichari

Thèse présentée pour l'obtention du grade de
*D*octorat en philosophie, Ph.D.
en Télécommunications

Jury d'évaluation

Examinateur externe	Prof. Soufian Ben Amor Université de Versailles St-Quentin-en-Yvelines
Examinateur externe	Dr. Hatim Chergui i2CAT Foundation-Research and Innovation Centre
Président du jury et examinateur interne	Prof. Tarek Djerafi INRS-EMT
Directeur de recherche	Prof. Sofiène Affes INRS-EMT

REMERCIEMENTS

Tout d'abord, je tiens à exprimer ma profonde gratitude envers les éminents membres du jury pour leur engagement et leur expertise lors de l'évaluation de ce travail.

Ensuite, Je saisis cette opportunité pour exprimer toute l'étendue de mon respect profond et de ma gratitude sincère envers mon superviseur émérite, le Professeur Sofiène Affes. Sa guidance éclairée, sa perspicacité aiguisée et son soutien permanent ont revêtu un rôle d'une importance cruciale dans la réalisation de cette thèse.

Je désire également exprimer ma profonde appréciation envers Dr. Oussama Ben Smida, dont les conseils stimulants et les discussions éclairées ont grandement enrichi cette recherche. Mes remerciements s'étendent également à mes collègues et amis proches, dont les échanges fructueux et le soutien moral ont contribué à enrichir cette expérience académique.

Enfin, Mes vifs remerciements à ma famille pour leur amour inconditionnel, leur soutien continu et leur compréhension inestimable à chaque étape de ce projet.

En somme, cette thèse est le fruit d'un effort collectif auquel je suis profondément reconnaissante envers chacun ayant joué un rôle dans sa réalisation.

RÉSUMÉ

La prochaine génération de réseaux cellulaires promet des améliorations considérables en termes de débits et de réduction de la latence, ouvrant la voie à une variété de nouveaux services. Cependant, cette avancée s'accompagne d'un défi majeur lié à l'efficacité énergétique, car le déploiement d'un grand nombre de stations de base (BS) devient indispensable pour assurer une couverture étendue et une efficacité spectrale accrue. Bien que les petites stations de base (SBS), telles que les picocellules, les microcellules et les femtocellules, soient conçues pour avoir des échelles individuelles plus petites et une consommation d'énergie moindre, l'utilisation combinée de ces stations suscite des préoccupations environnementales et économiques notables.

Face à la crise énergétique imminente, diverses méthodes sont proposées pour améliorer l'efficacité énergétique des futurs réseaux cellulaires, telles que l'optimisation de la gestion des ressources radio, les ajustements de la configuration des cellules, la mise en place de réseaux hétérogènes et l'utilisation de technologies de radio cognitive. La stratégie de commutation des stations de base (ON/OFF) émerge comme une solution pour accroître l'efficacité du réseau, mais elle présente des défis, en particulier dans les systèmes 5G avec des techniques innovantes de la couche physique et une architecture de réseau hétérogène. Les stratégies de commutation, qu'elles soient hors ligne ou en ligne, exigent une attention particulière en raison de leur impact sur la continuité du service et les opérations potentiellement coûteuses en énergie.

Malgré les progrès dans la conception des stratégies d'extinction des stations de base (BS), il persiste un besoin constant d'explorer divers critères pour des performances réseau respectueuses de l'environnement et efficaces.

Cette recherche se concentre sur les stratégies de commutation ON/OFF des stations de base (BS) dans les réseaux hétérogènes (HetNet), mettant l'accent sur la nécessité d'estimations précises des services futurs à partir d'analyses approfondies et d'outils d'estimation avancés. L'intégration de techniques d'apprentissage automatique (ML), en particulier d'apprentissage multimodal profond (DML), facilite la prédiction précise du trafic et la gestion adaptative des ressources du réseau. Différentes approches sont développées et testées pour la commutation en ligne des stations de base. L'algorithme de force brute (BF) établit une référence de performance pour l'activation et la désactivation des stations de base, en incorporant une fonction d'utilité pour un équilibre délicat entre la consommation d'énergie et les gains de débit. En conséquence, notre recherche nous permet d'atteindre des niveaux d'efficacité énergétique comparables à ceux de l'algorithme BF, tout en réduisant considérablement les coûts computationnels, à la fois en termes de temps et de ressources. Un avantage notable de ce modèle est sa capacité à économiser du temps, alignant ainsi les résultats d'optimisation avec les environnements réels et facilitant des ajustements rapides des cellules mobiles en réponse aux changements dynamiques des utilisateurs.

Mots-clés Réseau hétérogène, Efficacité énergétique, Commutation Marche/Arrêt, Consommation d'énergie, Apprentissage automatique, Apprentissage en profondeur, Algorithme Brute-Force

ABSTRACT

The forthcoming mobile generation shows promise with its expected improvements in bit rates and latency reduction, paving the way for a range of new services. Nevertheless, a significant issue arises concerning energy efficiency, as deploying a substantial number of base stations (BSs) becomes necessary for extended coverage and enhanced spectral efficiency. While small base stations (SBSs) like picocells, microcells, and femtocells are projected to have individually smaller scales and lower power consumption, the combined energy usage raises noteworthy environmental and economic concerns.

To address the impending energy crisis, various methods propose enhancing the energy efficiency of future mobile networks, including radio resource management optimization, cell configuration adjustments, heterogeneous network setups, and cognitive radio technologies. Base Station Switching (ON/OFF) emerges as a strategy for enhancing network efficiency but poses challenges, especially in 5G systems with innovative physical layer techniques and a heterogeneous network architecture. Switching strategies, offline or online, demand careful consideration due to their impact on service continuity and potential energy-costly operations.

Despite progress in designing Base Station (BS) switch-off strategies, there's a continuous need to explore diverse criteria for environmentally friendly and efficient network performance. Addressing the combinatorial optimization problem introduces complexities in network management and optimization strategies.

This research focuses on HetNet switch-off strategies, emphasizing the need for accurate future service estimations from comprehensive analyses and advanced estimation tools. The integration of machine learning (ML), particularly deep multi-modal learning (DML), facilitates precise traffic prediction and adaptive network resource management. Distinct approaches are developed and tested for online switching of base stations. The brute force (BF) algorithm establishes a performance baseline for base station activation and deactivation, incorporating a utility function for a delicate balance between energy consumption and throughput gains. Deep multi-modal learning (DML) trains two neural networks, predicting non-critical base stations and forecasting user-base station associations. As a result, our research enables us to reach energy efficiency levels comparable to those of the BF algorithm, while substantially cutting down on computational costs, both in terms of time and resources. A notable advantage of this model is its capacity to save time, aligning seamlessly with a key research objective : customizing optimization outcomes to real-world environments and facilitating prompt adjustments of mobile cells in response to dynamic user changes.

Keywords Heterogeneous network,energy efficiency, On/Off switching, power consumption, machine learning, deep learning, Brute-Force

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

3GPP	The Third Generation Partnership Project
5G	Fifth Generation Cellular Networks
6G	Sixth Generation Cellular Networks
AI	Artificial Intelligence
ARIMA	Autoregressive integrated moving average
BBU	Baseband Unit
BS	Base Station
CAPEX	Capital expenditure
CLA	Cell-Layout Adaptation
CO_2	Carbon dioxide
CoMP	Coordinated MultiPoint transmission
CR	Cognitive Radio
CRAN	Cloud-Radio Access Network
CRE	Cell Range Expansion
DL	Deep Learning
DML	Deep Multimodal Learning
DQL	Deep Q-Learning
DTX	Discontinuous Transmission
EE	Energy efficiency
eMBB	Enhanced mobile broadband
ENDC	E-UTRAN New Radio – Dual Connectivity
ETSI	European Telecommunications Standards Institute
GRU	Gated Recurrent Units
HetNets	Heterogeneous Networks
ICT	Information and Communications Technologies
IMT	International Mobile Telecommunications
ITU	International Telecommunication Union
KNN	k-nearest neighbours
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error

MBS	Macro Base Stations
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
mMTC	Massive machine-type communications
mmWave	millimeter-wave
MU-MIMO	Multi-User MIMO
NFV	Network Function Virtualization
NN	Neural Networks
NR	New Radio
OPEX	Operating expenditure
OTN	Optical Transport Network
QoS	Quality of Service
RL	Reinforcement learning
RNNs	Recurrent neural networks
RRH	Remote Radio Head
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
SARIMA	Seasonal autoregressive integrated moving average
SBS	Small Base Stations
SDN	Software Defined Network
SE	Spectral Efficiency
SINR	Signal to Interference Plus Noise Ratio
SL	Supervised learning
SMS	Short Message Service
SVR	Support Vector Regression
SWIPT	Simultaneous Wireless Information and Power Transfer
UDN	Ultra-Dense Networks
UE	User Equipment
URLLC	Ultra-reliable low latency communication
USL	Unsupervised learning

SOMMAIRE RÉCAPITULATIF

L'évolution des réseaux sans fil a été une aventure fascinante, et chaque nouvelle génération de réseau a apporté des améliorations significatives. Revenons à l'époque de la 1G, qui a posé les bases de la communication en utilisant une technologie analogique qui prend en charge uniquement la voix. Ensuite, la 2G est apparue et a fait un grand pas dans le monde de la communication numérique, permettant l'introduction de nouveaux services tels que les SMS (service de messagerie courte) et 'Roaming'. En 2000, l'UIT (Union internationale des télécommunications) a publié les Télécommunications mobiles internationales (TMI)-2000, fournissant un cadre mondial unifié pour la technologie 3G. Ce cadre englobait des aspects cruciaux tels que l'allocation du spectre de fréquences et les critères de performance. Pendant cette période, l'introduction des smartphones a révolutionné l'accessibilité à la navigation sur le web et aux services de messagerie électronique. En offrant un accès à Internet et des capacités d'appel vocal, la 3G a changé la donne et a modifié notre façon de rester connectés. À mesure que de nouveaux scénarios émergeaient, nécessitant une utilisation accrue des données, la nécessité de technologies innovantes est devenue évidente. L'arrivée de la 4G en 2010, sous la bannière de l'IMT-Advanced, a propulsé l'internet mobile vers de nouveaux sommets en introduisant des idées révolutionnaires telles que l'adoption d'une structure "tout IP", permettant le streaming vidéo sans faille et offrant des vitesses de bande passante mobile nettement plus rapides.

Actuellement, toute l'attention est portée sur la 5G, caractérisée par une vitesse et une réactivité exceptionnelles. Elle offre une connectivité plus rapide, une fiabilité accrue et une densité de connexion plus élevée, associées à une latence réduite, ouvrant la voie à des potentiels tels que les appareils intelligents et les véhicules autonomes. Cependant, le voyage ne s'arrête pas là. L'horizon du 6G se profile, apportant avec lui la promesse de vitesses encore plus étonnantes et d'activités en ligne presque instantanées. Au-delà de cela, la convergence du 6G avec l'intelligence artificielle (IA) est anticipée, une fusion qui pourrait révolutionner les industries et redéfinir notre manière d'interagir avec la technologie dans notre vie quotidienne. Cette perspective est incroyablement excitante, car elle offre le potentiel de remodeler complètement notre relation avec la connectivité.

Au cours des 50 dernières années, nous avons été témoins d'une évolution rapide des technologies sans fil, et elles deviennent une part de plus en plus importante de notre vie quotidienne (11). On s'attend à ce que leur importance ne cesse de croître, conduisant à de nouveaux types de services et de systèmes interconnectés. Alors que nous sommes en train de déployer la 5e génération (5G) des réseaux mobiles, des chercheurs du monde entier ont également commencé à examiner ce que les réseaux de la potentielle 6e génération (6G) pourraient offrir.

La révolution de la 5G se caractérise par son caractère distinctif. Il ne s'agit pas seulement d'une augmentation de la vitesse de nos téléphones, il y a une toute nouvelle manière de voir les choses. Un grand axe de recherche actuel porte sur l'efficacité énergétique, qui est devenue un facteur majeur dans la conception de la prochaine génération de réseaux cellulaires. Contrairement aux générations précédentes de réseaux, les réseaux 5G sont confrontés au défi de fournir simultanément une multitude de services, chacun ayant des exigences diverses en matière de qualité de service.

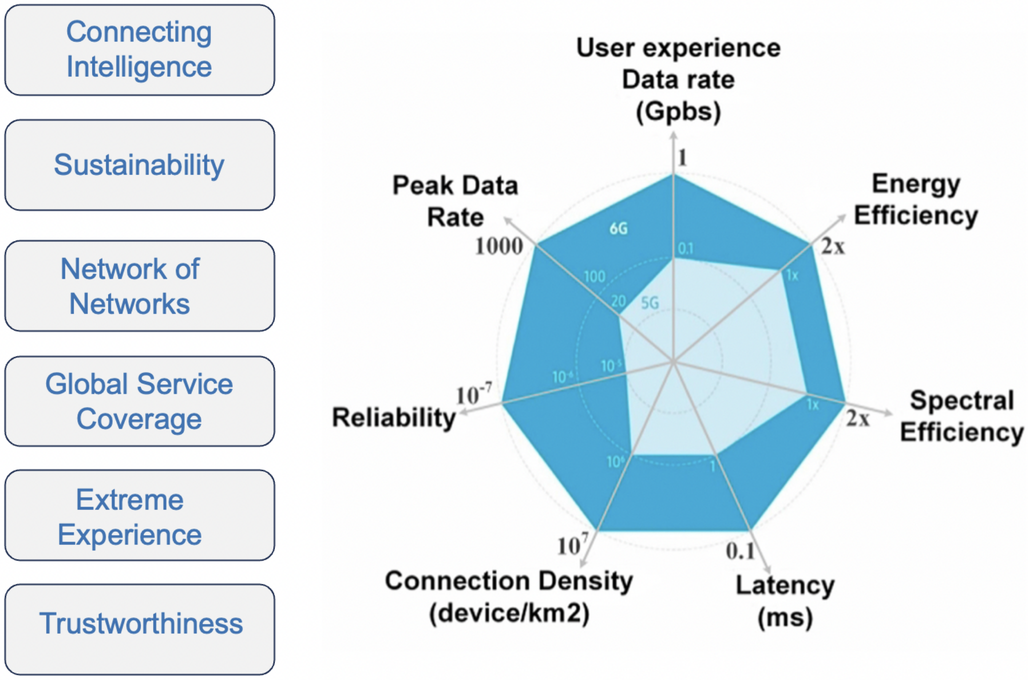


FIGURE 1 : Améliorations clés de la 6G par rapport à la 5G.

Lorsque nous le comparons à la capacité actuelle de la 5G, la 6G est censée apporter des débits de données plus élevés, une sécurité améliorée et une latence réduite. La vitesse projetée pour la 6G est estimée entre 1 et 10 Tbps. Sa fréquence dépassera celle de toutes les générations précédentes. Les fréquences augmentent généralement avec l'avancée de la technologie, et l'introduction de la fréquence térahertz (THz) indique des taux de transmission élevés. Avec la 6G, la latence est estimée entre 10 et 100 millisecondes, tandis que la densité de connectivité devrait être d'environ 10 appareils par kilomètre carré. La capacité de trafic devrait atteindre environ 1 Gb/s par mètre carré.

De plus, la 6G vise à améliorer certains indicateurs clés de performance (KPI), en poussant le pourcentage de couverture à 99%, la fiabilité à 99,9999%, en affinant la précision de positionnement du niveau du mètre à celui du centimètre, et en améliorant la sensibilité du récepteur à mieux que -130 dBm (204). L'efficacité spectrale et énergétique connaîtra des améliorations exponentielles par rapport à la 5G. L'un des aspects les plus excitants de la 6G est sa promesse de fournir une connectivité sans limites. Elle est conçue comme un réseau de communication complet intégrant divers systèmes, y compris la communication, la mesure, le stockage, l'informatique, le contrôle, le GPS, le radar, l'imagerie et la navigation (99). Les caractéristiques fondamentales de la 6G sont résumées dans la Figure 1 (23).

L'importance de l'efficacité énergétique dans le contexte des futurs réseaux sans fil ne peut être surestimée. Au-delà des défis techniques, des considérations économiques, opérationnelles et environnementales soulignent la nécessité de solutions durables et économes en énergie. La montée en puissance du nombre de stations de base et d'appareils connectés accentue l'urgence de passer d'une focalisation unique sur l'optimisation du débit à une approche globale qui privilégie l'efficacité énergétique.

D'un point de vue économique, les coûts opérationnels liés à la consommation d'énergie dans les réseaux sans fil sont substantiels. Une utilisation efficiente de l'énergie impacte directement la viabilité économique des opérateurs de réseau et des fournisseurs de services. Sur le plan opérationnel, l'optimisation de la consommation d'énergie s'aligne sur l'objectif de créer des réseaux sans fil résilients et durables capables de répondre aux demandes futures.

Prévenir la crise énergétique :

Avec la prochaine augmentation du trafic pour les systèmes cellulaires de nouvelle génération, il est urgent d'augmenter la capacité du réseau, nécessitant l'installation de plus de stations de base (BS). Ainsi, la prolifération de petites cellules émerge comme une solution prometteuse pour répondre aux exigences des systèmes sans fil 5G en termes de capacité réseau et de débit. L'intégration de petites cellules de différentes tailles aux côtés des stations de base macro crée un réseau hétérogène (HetNet) qui offre des performances élevées et une qualité de service pour gérer la hausse prévue du trafic. Cependant, l'augmentation du nombre de composants du réseau entraîne une hausse significative de la consommation d'énergie. Malgré l'efficacité énergétique inhérente des petites cellules, leur déploiement généralisé amplifie à la fois la consommation d'énergie du réseau d'accès radio et son empreinte carbone.

La consommation élevée d'énergie de l'industrie des Technologies de l'Information et de la Communication (TIC) entraîne environ 2% à 3% des émissions mondiales de dioxyde de carbone, représentant environ 25% de toutes les émissions (66). Par conséquent, pour réduire de manière significative les émissions mondiales actuelles de dioxyde de carbone dans les réseaux sans fil de prochaine génération, l'objectif est que les réseaux sans fil 5G fonctionnent en tant que réseaux respectueux de l'environnement avec des émissions de dioxyde de carbone notablement faibles. Cependant, les conceptions des réseaux sans fil cellulaires conventionnels ont principalement privilégié un débit utilisateur élevé et une grande capacité, avec une considération minimale pour l'efficacité énergétique ou la consommation d'énergie.

Par conséquent, des recherches approfondies sont menées pour minimiser les opérations des stations de base (BS) et améliorer l'efficacité énergétique globale du réseau, sous l'impulsion de considérations économiques, opérationnelles et environnementales (56), (128). En conséquence, l'efficacité énergétique est devenue un facteur crucial pour les performances des futurs réseaux 5G. Avec la perspective de millions de stations de base supplémentaires et de milliards d'appareils connectés, l'importance de la conception et de l'exploitation de systèmes économes en énergie a considérablement augmenté. Elle représente désormais un aspect fondamental dans le développement des réseaux de communication, marquant un changement d'accent de l'optimisation du débit vers la priorisation de l'optimisation de l'efficacité énergétique.

Pour faire face à l'imminente crise énergétique, la priorité donnée à l'efficacité énergétique est devenue un axe central dans la conception et l'exploitation des futurs réseaux mobiles. Diverses approches variées et complémentaires ont été suggérées pour accroître l'efficacité énergétique de ces réseaux. Ces méthodes incluent l'optimisation de la gestion des ressources radio, l'adaptation des configurations cellulaires, l'introduction de déploiements de réseaux hétérogènes et la mise en œuvre de technologies de radio cognitive, entre autres.

Motivations :

Les réseaux cellulaires ont eu un impact profond sur notre vie quotidienne, et la cinquième génération (5G) de technologie radio s'apprête à apporter encore plus de changements transformateurs. Elle promet de permettre des niveaux inédits d'automatisation et d'innovation dans divers secteurs en raison de sa capacité améliorée, de sa connectivité étendue et de communications à faible latence incroyablement fiables. La 5G est un réseau complexe qui prend en charge une large gamme de services grâce à plusieurs technologies clés. Celles-ci comprennent la virtualisation, qui permet une gestion de réseau plus flexible et efficace, la softwarisation, qui rend les fonctions réseau plus adaptables, de nouveaux Réseaux d'Accès Radio (RAN) pour une connectivité améliorée, et des stratégies novatrices de transport des données pour un meilleur acheminement. Ces technologies travaillent ensemble pour fournir une faible latence, un transfert de données à haute vitesse et des connexions fiables.

Cependant, alors que nous aspirons à une plus grande capacité réseau, une couverture plus étendue et une augmentation du trafic de données, nous sommes également confrontés au défi de la hausse de la consommation d'énergie. Cela n'est pas durable, tant d'un point de vue environnemental que commercial. Reconnaissant cela, il y a un effort mondial pour rendre nos réseaux plus économes en énergie, motivé par des préoccupations économiques et environnementales.

Les projections indiquent que le nombre d'appareils connectés pourrait potentiellement atteindre 100 milliards d'ici 2030(152), anticipant une augmentation significative du trafic de données, estimé à croître jusqu'à 1 000 fois plus qu'en 2018 pour la 4G(84). Cette augmentation substantielle de l'utilisation des smartphones, des dispositifs portables et des objets connectés (IoT) représente un défi notable en termes de fourniture de vitesses de données rapides, d'une couverture étendue et d'une latence minimale. De plus, avec chaque nouvelle génération de technologie sans fil, nous avons observé une augmentation de la consommation d'énergie due à l'ajout de matériel pour prendre en charge les applications émergentes et les exigences évolutives. La tendance suggère que la 5G continuera cette trajectoire, augmentant considérablement la consommation d'énergie en comparaison avec sa version précédente, la 4G. La nécessité d'accommoder des débits de données élevés et une multitude d'appareils connectés rend ces réseaux plus gourmands en énergie. En fait, on estime que la consommation d'énergie de la 5G pourrait être quatre fois supérieure à celle de la 4G(77). Par conséquent, la recherche de l'efficacité énergétique émerge comme une préoccupation cruciale dans le contexte de la 5G, la distinguant des générations précédentes.

Pour illustrer l'ampleur et la signification de l'impact de l'essor de la 5G, il est intéressant de souligner que le rapport SMART 2020 de l'Union Internationale des Télécommunications (UIT) (89) a estimé que les communications mobiles seules ont contribué à environ 2 135 millions de tonnes d'émissions de CO2e en 2018. De plus, les prévisions de (87) indiquent que cet impact devrait augmenter considérablement au cours de l'ère de la 5G. À la fin de l'année 2020, on prévoyait que l'effet cumulatif sur l'ensemble du secteur des Technologies de l'Information et de la Communication (TIC) représenterait environ 15% des émissions mondiales totales de gaz à effet de serre.

Face à cette problématique, le Projet Partenarial de Troisième Génération (3GPP) a introduit la spécification New Radio (NR), qui dote la prochaine génération de réseaux d'outils pour réduire significativement la consommation d'énergie et les émissions de gaz à effet de serre. Cela s'inscrit dans le cadre des objectifs plus larges de durabilité du secteur des Technologies de l'Information et de la Communication (TIC), contribuant à un avenir plus responsable sur le plan environnemental et plus efficient.

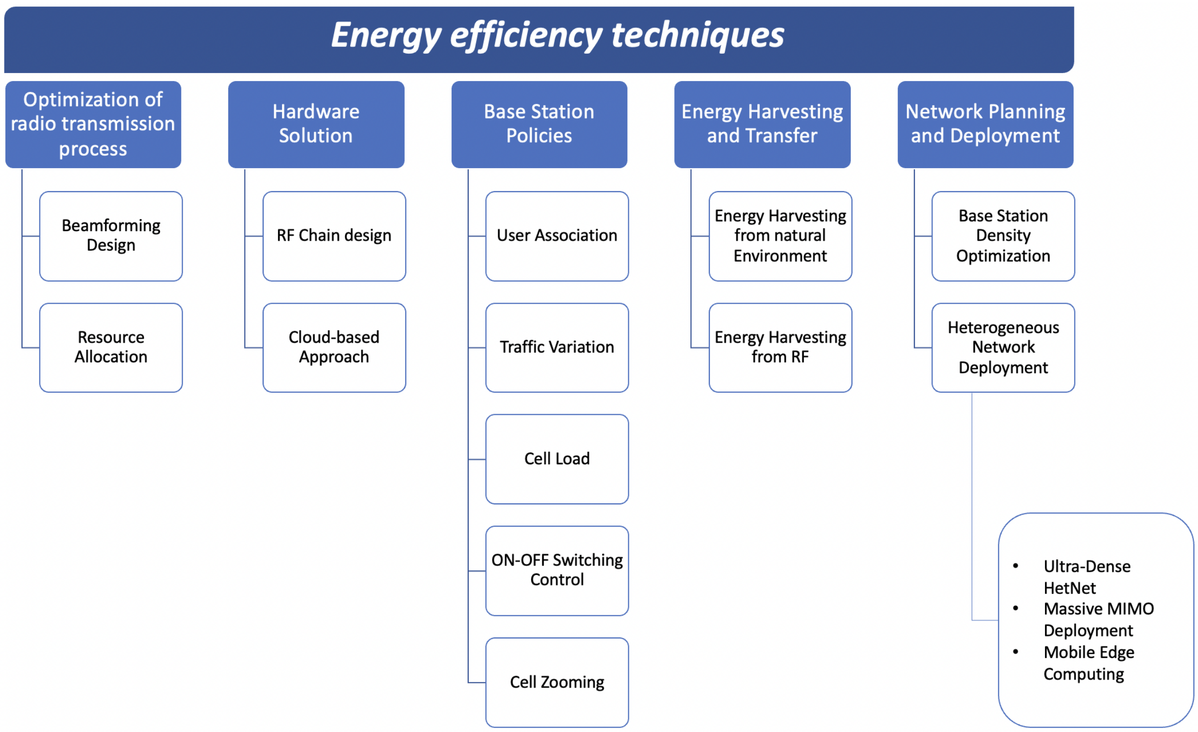


FIGURE 2 : Stratégies d'Optimisation Énergétique dans les Réseaux Mobiles Sans Fil

La réduction de la consommation d'énergie dans les réseaux cellulaires hétérogènes a récemment suscité un intérêt croissant. La littérature a présenté plusieurs méthodes pour concevoir des réseaux économes en énergie dans le contexte des opérations sans fil mobiles (comme indiqué dans la figure 2). Bon nombre de ces méthodes se concentrent sur l'amélioration de l'efficacité énergétique (EE) des stations de base (BS), car les unités BS sont responsables d'une part substantielle, allant de 60% à 80% (90), de la consommation totale d'énergie dans les réseaux mobiles sans fil. Par conséquent, plusieurs études ont été menées pour améliorer l'efficacité énergétique du réseau. Ainsi, les algorithmes d'activation/désactivation des BS figurent parmi les solutions d'économie d'énergie les plus puissantes. Ces algorithmes éteignent les parties inutiles du réseau (c'est-à-dire, les BS) et déplacent les utilisateurs vers les BS voisines pendant les périodes creuses. Ils activent également le nombre approprié de BS selon les besoins. Par conséquent, les algorithmes d'activation/désactivation des BS seraient particulièrement avantageux pour les futurs réseaux de cinquième génération (5G) caractérisés par une densification extrême des BS, rendant le défi de l'efficacité énergétique encore plus complexe.

La commutation des stations de base (activation/désactivation), qui représente l'une des stratégies potentielles pour améliorer l'efficacité du réseau, offre des perspectives significatives mais pose des défis d'implémentation notables. La complexité découle du fait que la désactivation des stations de base entraîne l'arrêt complet des services dans une zone donnée, en particulier lorsqu'on considère un modèle de couverture à un seul niveau. Cependant, cet obstacle peut être surmonté grâce à l'introduction des réseaux 5G, qui sont prévus pour étendre l'infrastructure de l'interface radio selon diverses tailles et hiérarchies, formant des réseaux hétérogènes (HetNets). On prévoit que ce paradigme HetNet créera des réseaux superposés comprenant de grandes stations de base macro (MBS) couvrant de plus petites stations de base (SBS) sous-jacentes.

Organiser les réseaux hétérogènes de cette manière équilibre le terrain de jeu pour les algorithmes d'activation/désactivation des stations de base en augmentant les besoins énergétiques globaux du réseau, rendant la technologie essentielle, tout en facilitant simultanément sa mise en œuvre grâce à la couverture superposée. Avant ce changement de paradigme, pour qu'une station de base soit éteinte, elle devait posséder un mécanisme pour surveiller efficacement les demandes de service afin d'éviter les interruptions de couverture. En revanche, au sein des HetNets, la désactivation d'une SBS inactive entraîne principalement une perte de capacité potentielle, qui, lorsqu'elle est bien planifiée, peut contribuer à des économies d'énergie sans perturber les opérations régulières du réseau. À mesure que les HetNets sont adoptés plus largement, le défi persistant d'éviter des interruptions de service complètes dans une zone est résolu de manière efficace, permettant de supposer que la commutation des stations de base est réalisable uniquement dans les régions avec une couverture superposée.

Plusieurs stratégies d'activation/désactivation des stations de base (BS) (76) ont été proposées sous différentes perspectives de conception pour optimiser exclusivement les économies d'énergie ou d'autres compromis de performance liés à l'énergie, tels que des stratégies aléatoires, basées sur la charge ou la distance. De plus, des recherches ont été entreprises pour envisager une conception commune de la stratégie d'activation/désactivation des BS et d'autres stratégies, telles que l'association d'utilisateurs, les stratégies d'annulation des interférences en couche physique et l'allocation des ressources.

Les stratégies d'activation/désactivation des BS peuvent être mises en œuvre en se concentrant soit sur les petites stations de base (SBS), soit sur les grandes stations de base macro (MBS), soit sur les deux types de BS dans les HetNets. Cependant, la désactivation des MBS, qui est dérivée en minimisant la consommation d'énergie des BS, peut avoir un impact négatif significatif sur la couverture du réseau. En revanche, la stratégie de désactivation aléatoire des SBS est conçue en maximisant l'efficacité énergétique (EE) tout en respectant la contrainte de probabilité de couverture. Beaucoup de travaux ont été réalisés pour concevoir une stratégie spécifique d'activation/désactivation des BS dans les HetNets. Néanmoins, des améliorations et des défis restent à explorer en exploitant de manière appropriée la combinaison de différents critères pour obtenir des performances réseau plus écologiques et meilleures.

Il existe généralement deux approches pour activer-désactiver correctement un certain nombre de stations de base (BS) (174), (144) : les approches hors ligne et en ligne. L'approche hors ligne, relativement simple, permet la planification préalable des intervalles d'activation-désactivation. Cependant, elle présente un inconvénient majeur, à savoir qu'elle ne prend pas en compte la charge instantanée actuelle et, par conséquent, n'est pas robuste face à des événements imprévisibles (c'est-à-dire, des défaillances, des points chauds aléatoires, etc.), ce qui restreint son efficacité. D'autre part, l'approche en ligne considère exclusivement la charge réelle (c'est-à-dire, instantanée) pour décider de la nécessité de désactiver une BS ou non. Ainsi, elle peut gérer des événements imprévisibles en activant la partie appropriée du réseau (c'est-à-dire, le nombre de BS) pour faire face à une augmentation ou une diminution inattendue du trafic. Malgré son importance, cette approche est complexe car la décision doit être prise et exécutée en temps réel. En outre, elle peut entraîner un grand nombre d'opérations d'activation-désactivation avec des coûts énergétiques importants, ce qui peut non seulement augmenter considérablement le coût énergétique global du réseau, mais aussi endommager le matériel.

Certaines études se concentrent sur l'utilisation de l'optimisation mathématique (57) pour gérer l'activation et la désactivation des stations de base (BS) de manière à maintenir l'expérience utilisateur tout en réduisant simultanément la consommation d'énergie. Ces algorithmes éteignent les parties inutiles du réseau (c'est-à-dire, les BS) et déplacent les utilisateurs vers les BS voisines pendant les périodes creuses. Ils activent également le nombre approprié de BS selon les besoins (36). Par conséquent, les algorithmes d'activation/désactivation des BS seraient particulièrement avantageux pour les futurs réseaux de cinquième génération (5G) caractérisés par une densification extrême des BS, rendant le défi de l'efficacité énergétique encore plus complexe.

À cette fin, les recherche se sont focalisées sur les économies d'énergie pendant les périodes de faible charge réseau, permettant ainsi la mise en veille des stations. La plupart de ces structures tirent parti de la variation du trafic tout au long de la journée : lorsque la charge est réduite pendant une période significative, notamment la nuit, certaines stations sont éteintes tandis que celles restant opérationnelles prennent en charge leurs utilisateurs (27), (100). Diverses techniques ont été mises en œuvre dans cette démarche afin de garantir le maintien d'un service satisfaisant. Ces approches présentent des avantages notables en termes de consommation d'énergie et de débit binaire. Toutefois, il est crucial de prendre en considération que l'environnement cellulaire est dynamique, avec des utilisateurs changeant constamment de positions et de nombres. Ainsi, le facteur temps revêt une importance primordiale, mais ces méthodologies requièrent un investissement en temps et en ressources informatiques considérable.

De nos jours, les avancées en intelligence artificielle (IA) ont conduit à une adoption généralisée des techniques d'apprentissage automatique pour optimiser les communications sans fil. Ces techniques ont été intégrées dans les réseaux auto-organisés (SON), dans le but de rendre les tâches quotidiennes des opérateurs de réseau plus rationalisées et efficaces (141). En incorporant des outils basés sur l'IA, la progression du paradigme SON dans la 5G évolue vers une méthodologie proactive. Cette méthodologie tire parti de la vaste quantité de données disponibles et intègre des dimensions supplémentaires dérivées de la caractérisation de l'expérience et du comportement des utilisateurs finaux (60).

À cet égard, divers schémas écoénergétiques utilisant l'apprentissage automatique ont été développés. Ces approches présentent des avantages significatifs en termes de consommation d'énergie et de débit binaire (58). Les principales stratégies pour optimiser l'efficacité énergétique au niveau de la station de base incluent les modes de veille adaptatifs basés sur l'apprentissage par renforcement (50), les stratégies d'extinction basées sur le trafic (73), et le contrôle efficace de la puissance de transmission (180). Cependant, il est important de prendre en compte qu'une cellule est un environnement dynamique où les utilisateurs changent constamment de positions et de nombres. De plus, ces méthodes basées sur l'apprentissage automatique ne permettent pas d'obtenir des économies d'énergie au même degré que l'approche mathématique, et elles impliquent également un coût substantiel.

Optimiser la commutation des stations de base présente encore certains défis qui nécessitent une attention particulière. En réponse à ces défis, il existe une richesse d'activités académiques visant à explorer diverses mises en œuvre possibles. Beaucoup de ces mises en œuvre identifient divers facteurs contribuant à la complexité. Cependant, le problème principal lié à la commutation des stations de base réside dans le fait que les équipements actuels ne sont pas conçus pour accommoder fréquemment les changements de mode opérationnel et exigent une attention particulière pendant le processus de commutation. Cette limitation dissuade les fournisseurs de services d'adopter fréquemment cette approche. L'incapacité à éteindre les stations de base à la demande nécessite une estimation précise des besoins futurs en service pour une zone donnée, un problème complexe qui repose sur une compréhension approfondie de l'historique des services de la zone et l'utilisation d'outils d'estimation sophistiqués. De plus, le défi de résoudre le problème d'optimisation combinatoire, qui détermine la combinaison la plus efficace de stations de base à activer à un moment donné, devient plus complexe à mesure que le nombre de combinaisons augmente.

Face à ces scénarios où les méthodes d'optimisation nécessitent un temps et des ressources informatiques substantiels, et où les approches d'apprentissage automatique (ML) n'ont pas atteint un équilibre satisfaisant entre coût et économies, nous avons reconnu le potentiel de combiner les forces des deux domaines. Notre objectif est d'obtenir des améliorations significatives de l'efficacité énergétique tout en réduisant simultanément le temps et les exigences en calcul. Cela nous a motivés à développer une approche d'apprentissage profond centrée sur la résolution de problèmes d'optimisation dans les réseaux sans fil en tenant compte de différents critères et perspectives. Cependant, une conception adaptable capable de répondre efficacement aux changements des conditions du réseau entraînerait des économies d'énergie plus importantes et une amélioration des performances du réseau.

En explorant les pistes pour rendre les réseaux hétérogènes mobiles sans fil plus écoénergétiques, plusieurs angles ont été examinés, allant de l'optimisation du déploiement des stations de base (BS) à l'amélioration des transmissions radio, en passant par la gestion des schémas de veille des BS et l'utilisation de sources d'énergie renouvelable. Cependant, ces efforts ne sont pas sans difficultés. Concevoir des HetNets écoénergétiques nécessite de prendre en compte divers facteurs critiques, ce qui présente des défis pour notre recherche. Cela englobe la gestion des petites stations de base (SBS) distribuées de manière aléatoire, la prise de décisions optimales concernant les associations d'équipements d'utilisateur (UE), l'allocation dynamique des ressources adaptée à la fluctuation des motifs de trafic, tout en respectant l'équilibre entre l'efficacité énergétique et les normes de service de haute qualité (QoS).

Méthodologie de recherche :

Dans le cadre de cette recherche, nous avons élaboré et testé des approches distinctes pour aborder la question de la commutation des stations de base dans le contexte des réseaux hétérogènes écoénergétiques. Dans une première étape, l'algorithme de brute force (BF) a été utilisé dans le contexte de l'activation et de la désactivation des stations de base, tout en incorporant une fonction d'utilité pour équilibrer la consommation d'énergie et les gains de débit en vue d'assurer la qualité de service (QoS). En analysant de manière exhaustive diverses combinaisons d'états de stations de base, la stratégie de brute force établit une référence de performance qui aide à évaluer l'efficacité et offre la meilleure solution possible pour la commutation de cellules. L'introduction de la fonction d'utilité, qui suggère l'équilibre délicat entre l'efficacité énergétique et les performances du réseau, contribue au processus de prise de décision. Cette première étape valide non seulement la capacité de gérer les stations de base pour une QoS optimale, mais offre également une compréhension réelle de l'interaction entre les considérations énergétiques et le débit de données.

La deuxième étape utilise l'apprentissage profond multi-modal (DML) pour entraîner deux réseaux neuronaux (NN). Le premier réseau neuronal a la capacité d'assimiler une large gamme de paramètres d'entrée (y compris des facteurs tels que le nombre d'utilisateurs, la puissance reçue, la distance entre les stations de base ...) pour prédire les stations de base non critiques qui peuvent passer en mode veille. Le deuxième réseau est dédié à la prévision des associations utilisateur-station de base en fonction des sorties d'état des stations de base générées par le premier réseau neuronal. En intégrant ces prédictions, une solution globale est présentée pour optimiser l'efficacité énergétique dans les réseaux sans fil. Dans cette approche, diverses implémentations potentiellement utiles de réseaux neuronaux artificiels (ANN) sont testées, avec deux principaux paradigmes étant les réseaux neuronaux récurrents (RNN) tels que LSTM ou GRU, et les architectures de DML basées sur le modèle fusionné proposé. L'objectif est d'identifier le modèle optimal capable d'améliorer l'efficacité énergétique tout en préservant la qualité de service.

Commencer par une méthode de brute force (BF), également appelée algorithmes de recherche exhaustive (RE), dans la résolution de problèmes offre une stratégie fondamentale qui aide à comprendre les complexités d'un problème. En explorant méticuleusement toutes les configurations possibles, elle établit une référence claire pour évaluer les gains d'efficacité réalisables grâce aux algorithmes optimisés ultérieurs. Cette méthode initiale sert de base pratique, démontrant la faisabilité du processus de commutation et mettant potentiellement en lumière des complexités inattendues. De plus, elle facilite une compréhension claire du comportement du réseau et assure une compréhension solide des subtilités du problème avant de se lancer dans des optimisations complexes.

Initier le processus avec une approche de brute force dans le contexte de l'activation et de la désactivation des stations de base, tout en incorporant une fonction d'utilité pour équilibrer la consommation d'énergie et les gains de débit pour garantir la qualité de service (QoS), constitue une étape fondamentale vers une gestion efficace du réseau. En analysant de manière exhaustive différentes combinaisons d'états de stations de base, la stratégie de brute force établit une référence de performance qui aide à évaluer l'efficacité et offre la meilleure solution possible pour la commutation de cellules. L'introduction de la fonction d'utilité, qui suggère l'équilibre délicat entre l'efficacité énergétique et la performance du réseau, contribue au processus de prise de décision. Cette étape initiale valide non seulement la capacité de gérer les stations de base pour une QoS optimale, mais offre également une véritable compréhension de l'interaction entre les considérations énergétiques et le débit de données.

À cet égard, l'objectif est de trouver une partition qui regroupe les stations de base les plus utilisées et minimise la somme de la consommation d'énergie tout en maintenant la qualité de service. Notre métrique proposée (5.7) tient compte à la fois du coût énergétique (conventionnel ou écologique) et de la qualité de service (QoS), où le compromis entre l'expérience perçue par l'utilisateur (c'est-à-dire, la performance du réseau) et les coûts opérationnels (OPEX) encourus, ajustés en fonction du trafic utilisateur.

Cependant, avec un nombre croissant de stations de base, la complexité computationnelle augmente de manière exponentielle, rendant l'approche de brute force intensément computationnelle et impraticable. Dans ce contexte, les algorithmes d'apprentissage automatique interviennent en exploitant des données historiques et des modèles pour prédire des configurations optimales de stations de base. Ces algorithmes s'adaptent aux conditions évolutives du réseau, optimisant la QoS tout en minimisant la consommation d'énergie. Grâce à l'analyse prédictive, ils peuvent prévoir les motifs de trafic et les besoins énergétiques, améliorant ainsi la QoS tout en tenant compte des contraintes énergétiques. Ainsi, tandis qu'une approche de brute force clarifie les dynamiques fondamentales, les algorithmes d'apprentissage automatique offrent une extensibilité et une adaptabilité pour gérer la nature dynamique et intensive en données.

Dans notre cas, avant d'adopter notre approche basée sur l'apprentissage automatique pour la prédiction, nous avons initialement utilisé la méthode de la 'brute force' comme référence, en tenant compte de sa nature exhaustive et de sa capacité à fournir des solutions parfaites. La méthode de la 'brute force' est ainsi perçue comme un repère pratique, représentant la limite de la précision de la prédiction dans un contexte où une solution exhaustive est connue.

Toutefois, pour établir une connexion entre notre approche d'apprentissage profond et les fondements théoriques de l'estimation statistique, nous introduisons le concept de la borne inférieure de Cramér-Rao (CRLB). La CRLB représente une limite théorique sur la précision atteignable par tout estimateur non biaisé dans le domaine de l'estimation statistique. Alors que la méthode de 'brute force' agit en tant que repère pratique, la CRLB offre une perspective théorique sur la meilleure précision théoriquement possible dans un cadre statistique.

Ainsi, en considérant la CRLB comme un lien conceptuel, notre choix d'utiliser la méthode de la 'brute force' en tant que repère pratique pour évaluer les performances de notre approche d'apprentissage profond est justifié. Lorsque notre modèle de prédiction atteint une précision parfaite, il se rapproche des performances de la méthode de la 'brute force', renforçant la validité de notre approche par rapport aux limites théoriques définies par la CRLB.

La synergie entre l'analyse initiale par BF et l'apprentissage automatique ultérieur confère une prise de décision efficace et adaptable dans des environnements de télécommunications complexes et dynamiques.

Malgré l'offre d'une solution convergente optimale, l'approche de l'algorithme BF nécessite des ressources computationnelles et du temps excessifs. Par conséquent, nous croyons que l'utilisation de l'apprentissage profond (DL) représente une excellente résolution. Cela s'explique par le fait qu'un réseau neuronal compétent, formé sur des données générées par l'algorithme précédent, peut produire des résultats comparables. Néanmoins, on s'attend à ce que ce processus soit moins gourmand en ressources et plus rapide en termes de calcul.

Entreprendre une initiative pionnière visant à améliorer l'efficacité énergétique des réseaux sans fil, notre recherche s'efforce d'utiliser des techniques d'apprentissage profond pour entraîner une paire de réseaux neuronaux.

Le premier réseau a la capacité d'incorporer une large gamme de paramètres d'entrée (y compris des facteurs tels que le nombre d'utilisateurs, la puissance reçue, la distance par rapport à la station de base, les stations de base voisines, la charge des dispositifs, la localisation de l'émetteur, etc.) pour prédire les stations de base non critiques et générer par la suite des sorties d'état des stations de base. Cet algorithme fournit des évaluations binaires pour chaque station de base individuelle, catégorisant ainsi leur état opérationnel comme étant soit 'ACTIF' soit 'INACTIF'.

Le deuxième réseau utilise la sortie du premier réseau en tant que paramètre d'entrée pour fournir les associations d'utilisateurs aux stations de base, offrant ainsi une résolution complète pour améliorer l'efficacité énergétique au sein des réseaux HetNet. Cette approche harmonieuse produit une solution globale qui contribue considérablement à l'amplification de l'efficacité énergétique, aboutissant finalement à une réduction significative de la consommation d'énergie.

La représentation schématique de notre modèle d'apprentissage profond (DL) proposé est illustrée dans la figure.3 suivante.

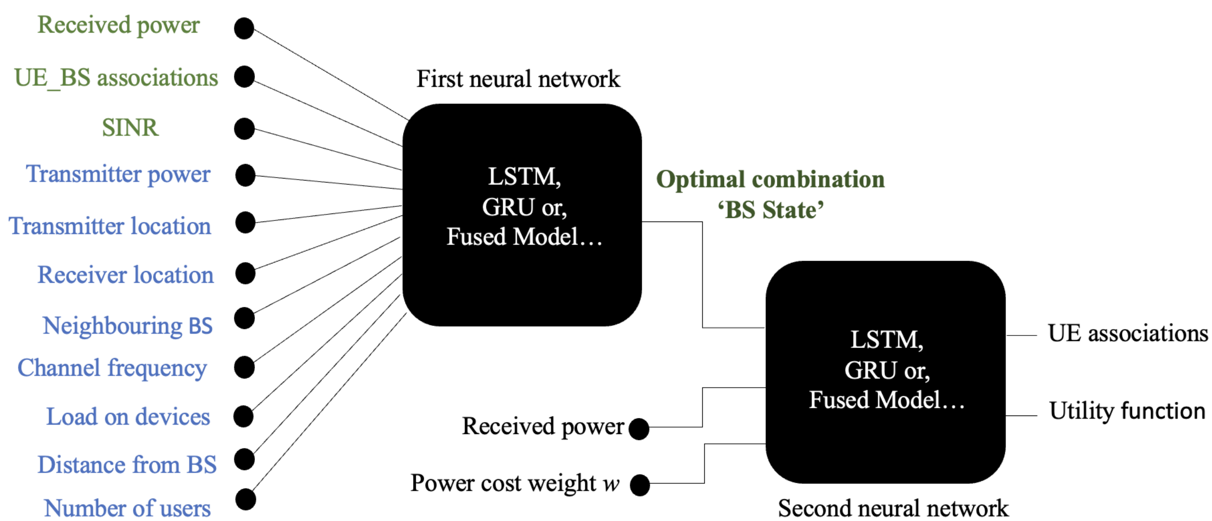


FIGURE 3 : Conception du Système

Pour faciliter l'entraînement de notre modèle, nous avons initialement généré un ensemble de données qui intégrait divers paramètres, tels que le nombre d'utilisateurs, la puissance reçue, la distance par rapport à la station de base, la station de base voisine, la puissance de l'émetteur, etc. Cet ensemble de données a été créé en exécutant plusieurs fois du code MATLAB pour la génération des paramètres. Par la suite, le réseau, tel que décrit précédemment, a été soumis à un entraînement en utilisant cet ensemble de données. Ces paramètres générés ont ensuite été catégorisés en combinaisons optimales de stations de base en utilisant différents modèles d'apprentissage profond.

En ce qui concerne la configuration de notre approche d'apprentissage profond, nos tests initiaux ont impliqué la mise en place de réseaux neuronaux selon trois modèles différents (LSTM, GRU et le modèle fusionné configuré). L'objectif était d'identifier le modèle optimal capable d'améliorer l'efficacité énergétique tout en préservant la qualité du service.

Le choix des architectures de réseau neuronal récurrent (RNN) telles que le Long Short-Term Memory (LSTM) et le Gated Recurrent Unit (GRU) pour prédire les stations hors service et améliorer l'efficacité énergétique dans les réseaux sans fil est délibéré et les distingue d'autres modèles d'apprentissage automatique. Conçus spécifiquement pour atténuer le problème du gradient qui s'annule, les LSTMs et les GRUs excellent dans le traitement de tâches séquentielles telles que la prédiction de séries temporelles. Les LSTMs exploitent une structure de cellule mémoire complexe pour capturer des dépendances prolongées, tandis que les GRUs utilisent une architecture plus simple avec des mécanismes de portes, ce qui peut conduire à un entraînement plus rapide et à de meilleures performances sur des tâches avec des dépendances plus courtes. La préférence pour les LSTMs et les GRUs par rapport à d'autres modèles d'apprentissage automatique découle de leur capacité innée à modéliser efficacement les dépendances temporelles, un facteur critique pour aborder la nature dynamique des données des réseaux sans fil. Ce choix délibéré souligne la décision stratégique d'utiliser des architectures RNN adaptées aux données séquentielles, les distinguant comme des solutions plus appropriées pour les complexités liées à la prédiction des stations hors service et à l'optimisation de l'efficacité énergétique dans les réseaux sans fil.

Pour améliorer les performances de notre système, un modèle multimodal d'apprentissage profond (DML) a été construit en s'inspirant de la méthodologie présentée dans l'article "Deep Multimodal Learning : Merging Sensory Data for Massive MIMO Channel Prediction" de Yang et al.(188). Bien que nous ayons adopté sélectivement une partie de leur modèle, nous avons adopté une approche distincte en formulant une architecture nouvelle. Ce modèle intègre une série de modèles LSTM avec des couches denses, incorporant des fonctions d'activation LeakyReLU, pour prédire efficacement l'état des stations de base. Notamment, notre approche exploite la puissance reçue ou bien SINR comme paramètre d'entrée pour les séquences LSTM, tandis que l'association d'utilisateurs sert d'entrée pour la couche Dense avec LeakyReLU. Cette fusion de flux de données divers permet à notre modèle de capturer des motifs complexes et de fournir des prédictions précises dans des scénarios réels complexes.

Composé de deux réseaux, le modèle fusionné intègre plusieurs couches LSTM (Long Short-Term Memory) dans le premier réseau, conçu pour analyser des données de séries temporelles et capturer des motifs temporels. Le deuxième réseau comprend plusieurs couches denses, chacune utilisant la fonction LeakyReLU, à l'exception de la couche de sortie, qui sert de fonction d'activation. Le réseau concaténé combine les sorties des deux modèles, qui comprennent à leur tour des couches denses avec la fonction LeakyReLU, intégrant efficacement ces modalités au niveau des décisions.

La couche dense, également appelée couche entièrement connectée, représente un type fondamental de couche de réseau neuronal. Chaque neurone de cette couche est connecté à chaque neurone des couches précédentes et suivantes. Cette couche se distingue par son utilisation de poids et de biais pour apprendre des motifs complexes dans les données lors du processus d'entraînement. Elle est couramment utilisée dans divers modèles d'apprentissage profond pour des tâches telles que la classification, la régression et l'apprentissage des caractéristiques.
Les couches LSTM sont spécifiquement conçues pour traiter des données de séries temporelles et capturer des dépendances temporelles, tandis que les couches denses avec des fonctions d'activation LeakyReLU sont des couches plus polyvalentes utilisées pour l'extraction de caractéristiques et la cartographie non linéaire dans les réseaux neuronaux.

La fonction d'activation LeakyReLU est une variation de l'unité linéaire rectifiée (ReLU) couramment intégrée dans les réseaux neuronaux. Elle fonctionne en appliquant une transformation non linéaire aux sorties des couches du réseau. Mathématiquement, la fonction LeakyReLU est définie comme $FLR(x) = \max[x, 0.2x]$, où x représente l'entrée de la fonction. En introduisant une pente légère (0.2) pour les valeurs négatives de x, elle permet à certaines informations de circuler même lorsque l'entrée est négative. Cela évite l'apparition du phénomène de "ReLU morte", où les neurones peuvent devenir inactifs, entravant l'apprentissage. L'implémentation de la fonction d'activation LeakyReLU, au lieu de la ReLU standard, vise à améliorer les performances et les capacités d'apprentissage du réseau neuronal.

En résumé, le modèle fusionné utilisant des couches LSTM et Dense, avec des couches denses suivies de la fonction LeakyReLU, sera utilisé pour prédire la combinaison des stations de base activées et désactivées en utilisant la puissance reçue et les associations d'utilisateurs en tant qu'entrées. Le modèle sera entraîné sur un ensemble de données comprenant la puissance reçue ou bien SINR et les associations d'utilisateurs en tant que caractéristiques d'entrée et les combinaisons de stations de base activées et désactivées en tant qu'étiquettes cibles. Les couches LSTM du modèle peuvent capturer les dépendances temporelles dans les données d'entrée, tandis que les couches Dense avec des fonctions d'activation LeakyReLU peuvent introduire une non-linéarité et apprendre des motifs complexes dans les données. En entraînant le modèle sur un ensemble de données suffisamment grand et diversifié, il peut apprendre à prédire les combinaisons de stations de base activées et désactivées en fonction des entrées fournies.

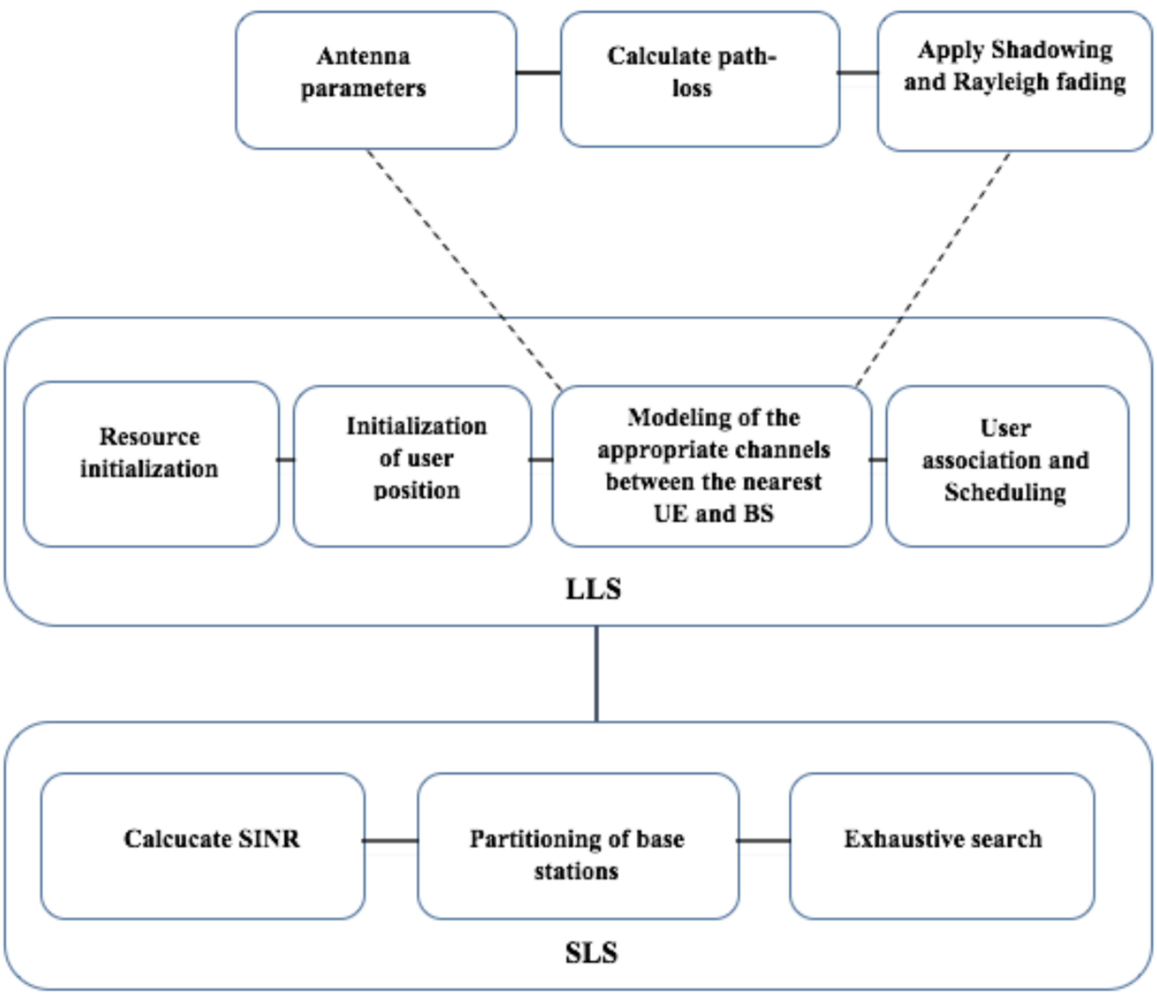


FIGURE 4 : Structure du simulateur

Pour évaluer les performances de nos algorithmes proposés, une plateforme de simulation a été développée, définissant la structure de notre travail. Nous avons opté pour un simulateur basé sur LTE, car notre intérêt se concentre principalement sur le signal de liaison descendante, et la 5G utilise le même signal que LTE. Comme notre réseau était hétérogène, nous avons déployé des stations de base micro et macro pour améliorer l'expérience des utilisateurs en termes de qualité de service. Ainsi, notre simulateur est conçu pour prendre en compte les deux environnements, à savoir les cellules macro et picocellulaires. En vue de la mise en œuvre et du déploiement de notre réseau, nous procéderons à une simulation à deux niveaux de notre HetNet local : une simulation de niveau liaison (LLS), qui concernait le canal radio entre l'antenne émettrice et l'antenne réceptrice où nous avons généré un modèle de canal couvrant la propagation des macrocellules et microcellules urbaines, ainsi que des simulations de niveau système (SLS) qui représentaient la fonctionnalité entre les stations de base et les utilisateurs finaux (BS et UE). À ce niveau, nous avons décrit l'allocation des ressources et l'association des utilisateurs, nous fournissant suffisamment de données (telles que le SINR et le débit) pour évaluer les performances de notre réseau. Ainsi, notre simulateur nous a fourni des indicateurs clés de performance nécessaires à notre approche centrée sur l'efficacité énergétique. La figure.4 illustre graphiquement la séquence de création de notre simulateur programmé. L'évaluation de la qualité du canal entre les stations de base environnantes et les utilisateurs nous guide dans la génération d'un canal à chaque TTI. (intervalle de transmission).

Les résultats démontrent des niveaux d'efficacité énergétique comparables entre l'algorithme de brute force et l'approche d'apprentissage multimodal profond. Cependant, cette dernière réduit significativement les coûts computationnels, atteignant l'objectif principal de la recherche d'adapter les résultats d'optimisation aux environnements réels. Cette réduction des coûts computationnels est attribuée à l'efficacité de l'apprentissage multimodal profond dans la prédiction des stations de base non critiques et des associations utilisateur-station. Les avantages et les inconvénients de chaque approche sont minutieusement analysés, fournissant des perspectives sur leur applicabilité dans différents scénarios de réseau. Les implications pratiques des résultats de la recherche offrent des perspectives sur la manière dont ils peuvent être appliqués dans des environnements réseau réels.

Conclusion

Les exigences croissantes en matière de trafic dans les réseaux cellulaires ont suscité la nécessité urgente de réévaluer l'impact environnemental de ces demandes en pleine expansion. La montée exponentielle de la consommation de données pose non seulement une menace à l'efficacité opérationnelle des réseaux de communication sans fil, mais contribue également de manière significative à l'émission de gaz à effet de serre, aggravant les défis environnementaux auxquels nous sommes confrontés. C'est dans ce contexte que l'impératif d'améliorer l'efficacité énergétique au sein de ces réseaux devient non seulement une nécessité financière, mais également une étape cruciale pour atténuer les conséquences écologiques de la connectivité moderne.

Du point de vue d'un opérateur, la quête de réduction de la consommation d'énergie dans les réseaux sans fil n'est pas uniquement motivée par des préoccupations environnementales. En effet, une réduction de la consommation d'énergie se traduit directement par des dépenses opérationnelles (OPEX) plus faibles. Cette double impérative de réduire à la fois les coûts financiers et environnementaux est un moteur clé pour les fournisseurs de services visant à créer des réseaux durables et économiques. Le programme Green Radio émerge comme un phare d'espoir dans ce paysage, aspirant à réaliser une réduction remarquable de cent fois de la consommation d'énergie par rapport aux conceptions actuelles des réseaux de communication sans fil.

Dans la poursuite de l'amélioration de l'efficacité énergétique, notre recherche marque une étape significative. En introduisant le concept de Deep Multimodal Learning (DML) dans le domaine des communications sans fil, nous avons atteint des niveaux d'efficacité comparables à l'algorithme bien établi de la Brute Force (BF). Ce qui distingue notre approche, cependant, est la réduction substantielle des coûts computationnels, tant en termes de temps que de ressources. Cette innovation ouvre de nouvelles perspectives pour trouver l'équilibre délicat entre l'efficacité opérationnelle, la responsabilité environnementale et la prudence financière dans le paysage dynamique des réseaux de communication sans fil.

Les résultats des simulations ont prouvé la puissance des architectures basées sur le DML, démontrant leur supériorité par rapport aux approches conventionnelles. La capacité à exploiter des données multimodales a révélé de nouvelles dimensions dans l'optimisation des réseaux de communication sans fil. Notamment, notre modèle proposé excelle dans le gain de temps, un facteur crucial dans notre effort pour adapter rapidement les résultats d'optimisation à des environnements réels. L'adaptabilité du modèle permet des ajustements rapides des cellules mobiles face aux changements dynamiques des utilisateurs, répondant à un besoin urgent dans le paysage en constante évolution des communications sans fil.

En regardant vers l'avenir, nos algorithmes proposés de commutation des stations de base (BS) off/on présentent une promesse particulière, surtout à mesure que nous anticipons des réseaux de prochaine génération caractérisés par une densification extrême des BSs. Le défi de l'efficacité énergétique prend une nouvelle couche de complexité, et nos approches suggérées se distinguent par leur flexibilité, leur adaptabilité, leur faible complexité et leur évolutivité. Importamment, elles peuvent intégrer facilement des critères d'optimisation supplémentaires, tels que les coûts d'investissement (CAPEX) et autres coûts opérationnels (OPEX), en ajustant simplement la métrique proposée.

En conclusion, notre recherche aborde non seulement le besoin pressant de réseaux de communication sans fil économes en énergie, mais jette également les bases de solutions innovantes qui concilient les préoccupations environnementales avec les exigences croissantes des systèmes de communication contemporains. Alors que nous naviguons dans les complexités d'un monde connecté, la synergie entre la rentabilité, la durabilité écologique et l'innovation technologique devient la boussole guidant l'évolution des communications sans fil vers un avenir plus durable.

1 Challenges of 5G Evolution to Future 6G

The evolution of wireless networks has been a fascinating adventure, and each new generation of network has brought significant improvements. Think back to the era of 1G, which established the groundwork for communication, using analog technology that support voice only. Then 2G appeared and took a big step into the world of digital communication enabling the introduction of new services like SMS (short message service) and roaming. In 2000, the ITU (International Telecommunication Union) released the International Mobile Telecommunications (IMT)-2000, providing a unified worldwide framework for 3G technology. This framework encompassed crucial aspects like frequency spectrum allocation and performance benchmarks. During this time, the advent of smartphones revolutionized the accessibility of web browsing and email services. By offering internet access and voice calling capabilities, 3G changed the game and changed the way we stay connected. As new scenarios emerged, requiring higher data usage, the necessity for innovative technologies became evident. The arrival of 4G in 2010, under the banner of IMT-Advanced, took mobile internet to new heights by introducing revolutionary ideas like adopting an "all IP" structure, enabling seamless video streaming and delivering notably faster mobile broadband speeds.
Currently, the spotlight is on 5G which is characterized by highest speed and responsiveness. It offers quicker connectivity, enhanced reliability, and greater connection density, coupled with reduced latency, paving the way for potentials such as smart devices and autonomous vehicles. Yet, the journey doesn't end here. The horizon of 6G is looming, bringing with it the promise of even more astounding speeds and nearly instant online activities. Beyond this, the convergence of 6G with artificial intelligence (AI) is anticipated, a fusion that could revolutionize industries and redefine how we interact with technology in our daily lives. This prospect is incredibly exciting, as it holds the potential to completely reshape our relationship with connectivity.

1.1 5G and 6G Systems Usage Scenarios

Over the last 50 years, we've witnessed a rapid evolution in wireless technologies, and they're becoming a bigger part of our daily lives (11). It's expected that their importance will only continue to grow, leading to new types of services and interconnected systems. As we're in the midst of rolling out the 5th Generation (5G) of mobile networks, researchers around the world have also begun looking into what the potential 6th Generation (6G) networks could offer.

The 5G revolution is characterized by its distinctiveness. It's not just about a speed boost for our phones, there's a whole new way of looking at things. One big focus now is on Energy Efficiency that's become a major factor in how we design the next generation of cell networks. Unlike the previous generations of networks, 5G networks face the challenge of concurrently delivering a multitude of services, each with diverse requirements for service quality. Currently, there is a

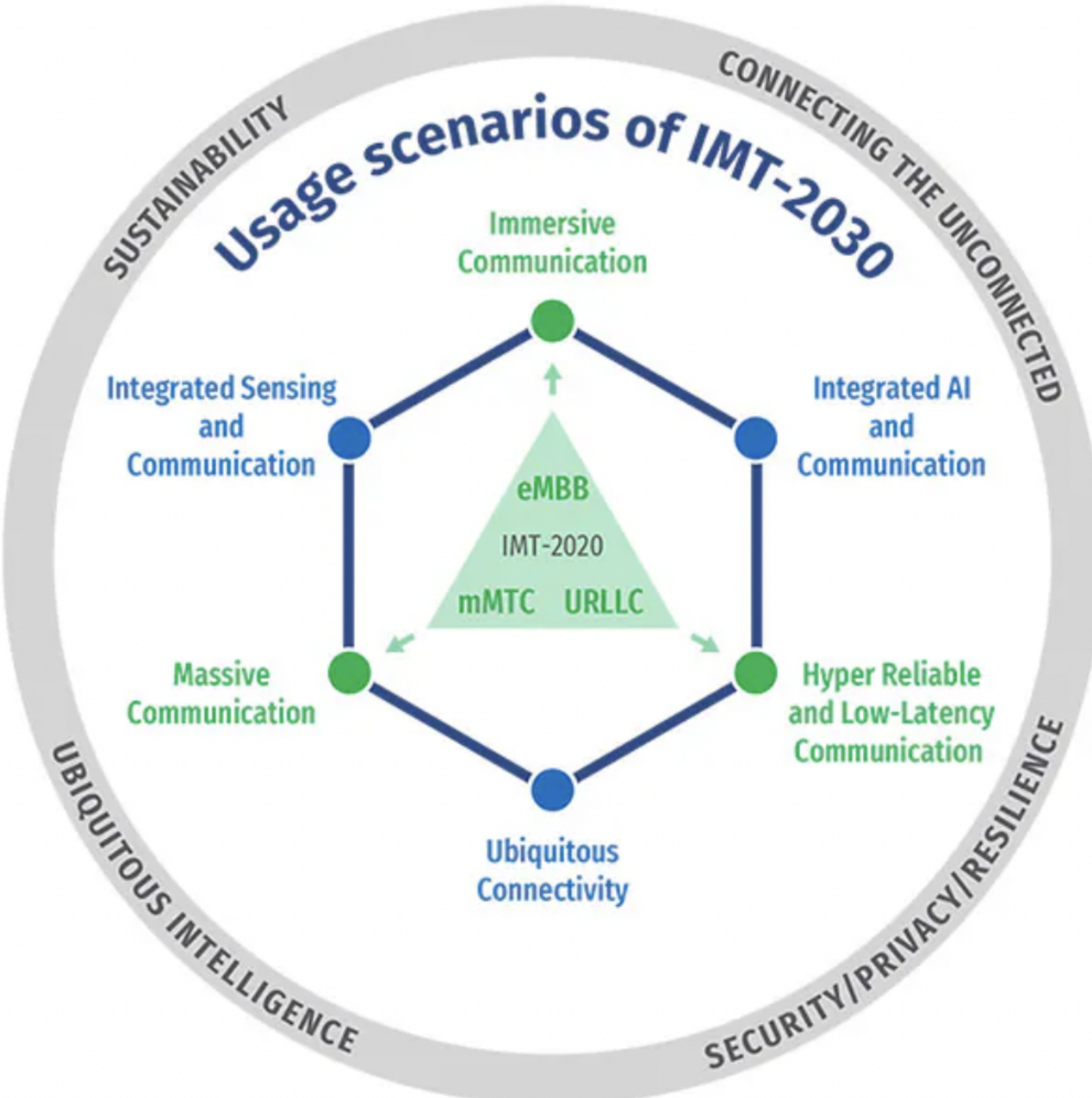


FIGURE 1.1 : Usage scenarios and general aspects of IMT-2030

widely accepted framework that classifies these services, as established by the International Telecommunication Union (ITU) and the International Mobile Telecommunications-2020 (IMT-2020). These services are classified into three distinct categories (148) :

Enhanced mobile broadband (eMBB) : serves people's needs for multimedia content, services, and data access. As demand rises, enhanced Mobile Broadband emerges with new applications and better performance. This involves different situations like wide-area coverage and hotspots, each with their own data capacity and user experience requirements.

Ultra-reliable low latency communication (URLLC) : This usage scenario will be highly significant in future applications that require both very high responsiveness and extremely reliable transmission.For instance, it might involve things like wirelessly controlling industrial manufacturing, performing medical surgeries remotely, handling the distribution of smart grids, ensuring safety in transportation, and more.

Massive machine-type communications (mMTC) : Facilitate extremely rapid connections capable of handling a significant number of connected devices, often seen in Internet of Things (IoT) scenarios.The automation of industrial processes through massive machine-type communications are examples of use cases that will be possible in future 5G systems. This application scenario requires a high quality of service in terms of energy efficiency, reliable connectivity, and strong reliability.

It's clear that the technical requirements for 6G applications can't be fulfilled by the existing 5G usage scenarios. So, researchers have started looking into how 6G could work by broadening the scope of the current usage scenarios. They're thinking about three new scenarios that could meet the needs of 6G use cases. These scenarios involve aspects that are common to both 5G and 6G, resulting in a comprehensive and holistic framework.

IMT-2030 (6G) encompasses six main usage scenarios, shown as a hexagon in Figure 1.1, extending from the IMT-2020 triangle (148). Surrounding this hexagon, you'll find a circle containing four overarching aspects : sustainability, ubiquitous intelligence, security/privacy/resilience, and connecting the unconnected, These serve as crucial design principles that are relevant to all usage scenarios.

1.2 Different Requirements of 6G Versus 5G

Figure 1.2 presents a comprehensive view of IMT-2030's capabilities, including nine enhanced aspects (such as peak data rate, spectrum efficiency, user experienced data rate, area traffic capacity, mobility, connection density, reliability, latency, and privacy/resilience/security) and six new dimensions (like coverage, sensing-related capabilities, positioning, AI-related capabilities, interoperability, and sustainability).

5G is on its way to widespread availability, setting the stage for the upcoming 6G. If we look at the evolution of generations, it's evident that internet speed and coverage gradually improve. The primary goal of 6G is to establish global coverage.AI applications will set 6G apart from previous generations.

When we compare it to the current capacity of 5G, 6G is expected to bring higher data rates, improved security, and reduced latency. The projected speed for 6G is estimated to range from 1 to 10 Tbps. Its frequency will surpass that of all previous generations. Frequencies generally increase as technology advances, and the introduction of Terahertz (THz) frequency indicates high transmission rates. With 6G, latency is estimated to be between 10 and 100 milliseconds, while the connectivity density is expected to be around 10 devices per square kilometer. Traffic capacity should be around 1 Gb/s per square meter. Moreover, 6G aims to enhance certain key performance

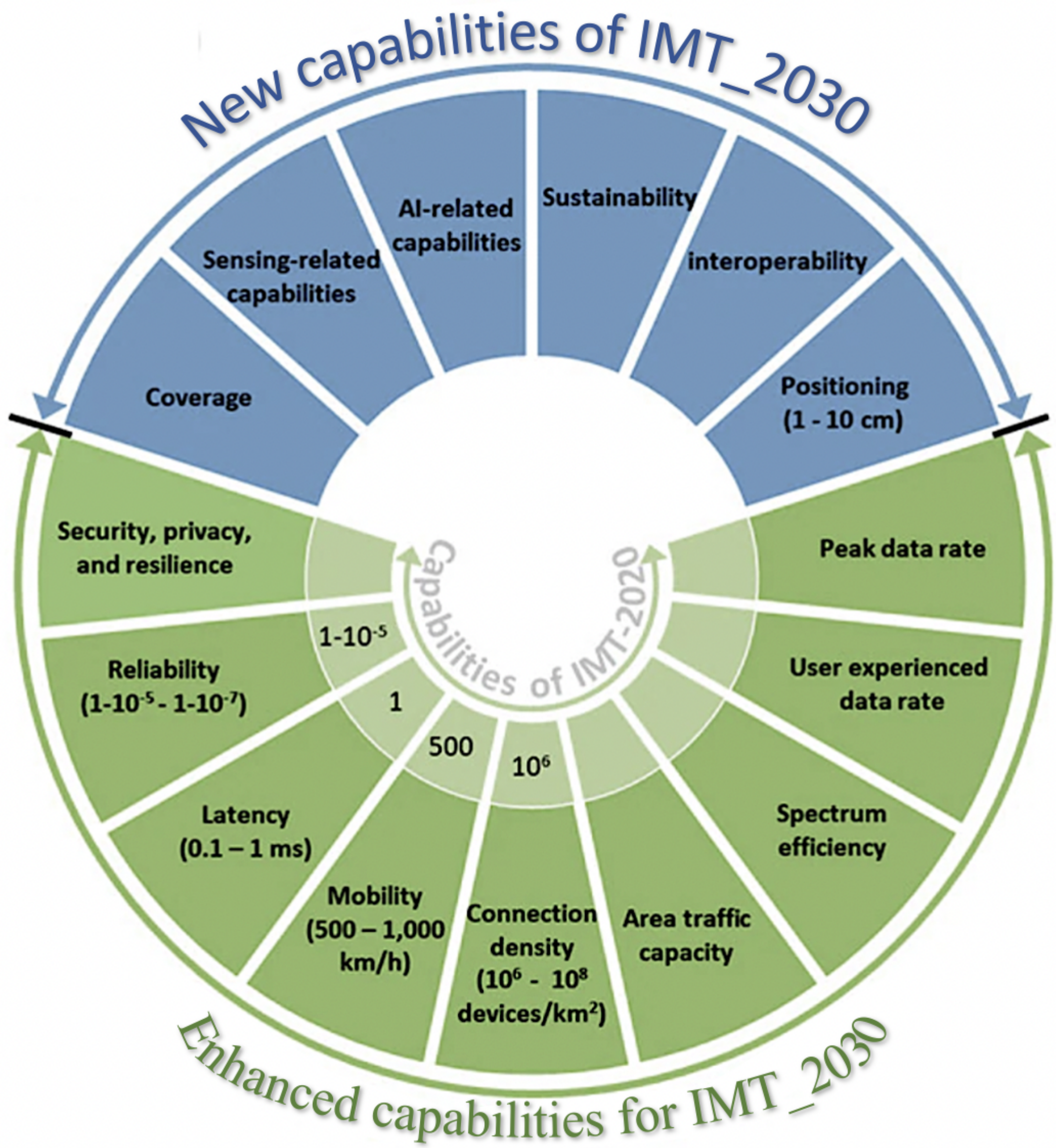


FIGURE 1.2 : Capabilities of IMT-2030

indicators (KPIs), pushing coverage percentage to 99%, reliability to 99.9999%, refining positioning accuracy from meter-level to centimeter-level, and improving receiver sensitivity to better than -130dBm (204). Spectrum and energy efficiency will see exponential improvements over 5G. One of the most exciting aspects of 6G is its promise of delivering limitless wireless connectivity. It's being designed as a comprehensive communication network accommodating various systems including communication, metering, storage, computing, control, GPS, radar, imaging, and navigation (99). The core characteristics of 6G are summarized in Figure 1.3 (23).

The most prominent challenges that future generations of 2030s networks must confront and lay the foundations for :

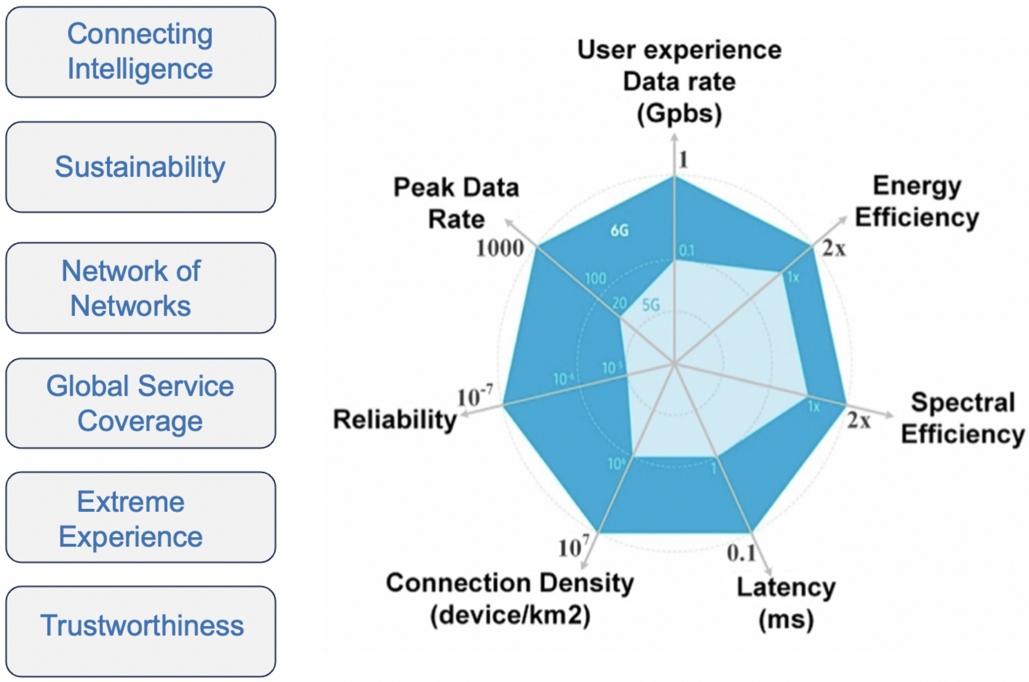


FIGURE 1.3 : Key enhancements of 6G versus 5G.

— **Connecting Intelligence :**The upcoming network generation needs to combine enhanced intelligence for more efficient data transfer while also providing the tools to use intelligent processes on a bigger scale across society.

— **Sustainability :**With technology's impact on the environment becomes more significant, the development of future communication systems like 6G focuses on minimizing their environmental footprint. 6G aims to make the most efficient use of resources across the network and improve energy efficiency. Additionally, 6G networks are anticipated to play a pivotal role in enhancing the sustainability of various sectors in the economy by providing the necessary digital tools to meet this objective.

— **Network of Networks :** 6G will have the task of accommodating various communication and computing resources that collaborate and link within a worldwide digital ecosystem. This ecosystem must bring together different scales, operating smoothly as a network of intelligent networks.Despite its complexity, this setup should make it easy and affordable to bring this technology to everyone, while providing energy and sustainability across all sectors of society.

— **Global Service Coverage :** As 6G networks combine different types of network technologies into a network of networks, they also have the task of ensuring worldwide coverage. The upcoming network generation must offer effective and reasonably priced digital access to connect individuals in remote or under-served areas. Furthermore, they're intended to support businesses and transportation by offering coverage over vast geographical areas.

— **Extreme Experience :**6G is predicted to bring several improvements, including faster speeds, higher capacity, shorter delays, and more accurate location and sensing services, all of which will outperform the capabilities of 5G networks.

— **Trustworthiness :** Given the growing significance of digital ecosystems in our daily lives, it's crucial for 6G networks to be robust and secure infrastructures. These networks must ensure the confidentiality and privacy of data communications. Establishing a trustworthy network is vital for enabling secure services within the democratic and sovereign societies.

1.3 AI and Machine Learning in the Optimization of 5G and 6G

To tackle the challenges that the next generation networks presents, we recognize the fundamental role of technologies like Artificial Intelligence (AI) and Machine Learning (ML). These technologies are expected to make a significant impact. While certain AI components were already in use in parts of the 5G network, we anticipate their widespread adoption and integration into the 6G networks. AI capabilities will gradually become an integral part of the entire 6G system, which represents a significant advancement. In particular, while AI-based communication design is expected to improve network performance, including improving data transmission efficiency, 6G networks are also expected to actively support and improve reliable AI/ML technologies on a broader level. This development will be advantageous for society, as it will provide interconnected, sustainable, and trustworthy intelligence. Basically, AI-driven communication will enhance network performance, and the 6G network itself will play a crucial role in making AI technologies more accessible and beneficial for everyone.

Artificial intelligence : AI, the technological trend of today, provides several beneficial features for 6G development. It's important to highlight that many researchers and industries have already embarked on the journey towards the next generation network, engaging in discussions about the technologies that will pave the way for 6G following the deployment of 5G. The role of AI in shaping the structure of 6G networks Holds great significance owing to its intelligent applications covering various domains like architecture, computing, storage, and more. The incorporation of AI into the 6G network involves the application of intelligent analysis techniques. An exemplary approach is descriptive analysis, which enhances our comprehension of historical data and network operations. Predictive analytic leverages data collected to make informed future predictions, while diagnostic analytic assumes a critical role in enhancing network security and enabling autonomous detection. Prescriptive analytic takes these predictions and provides multiple actionable insights(149). Moreover, AI facilitates the concept of closed-loop optimization within wireless 6G networks. In this context, user actions are guided by feedback from their devices. By maintaining end-to-end optimization at the system level, AI further elevates the efficiency of wireless communication systems. This intelligent layer introduces an array of possibilities, including self-optimization, self-learning, and comprehensive data collection (105). These examples provide just a glimpse of the extensive applications of AI technologies and the transformative innovations they hold for shaping the structure of 6G networks.

Machine Learning : ML, in the context of self-learning and decision-making, stands out as a technology of utmost importance. This AI-driven technology functions as a computational system that builds mathematical models by discerning system behaviors and characteristics. This enables machines to acquire knowledge without the need for explicit programming. ML algorithms draw upon data collected from diverse sources, using this data to refine algorithms across different levels, encompassing both the physical layer and higher layers (9). In the perspective of wireless communication, energy efficiency can sometimes be a concern. ML provides a solution by allowing devices to learn and enhance energy efficiency through adaptive decision-making and predictive capabilities. This capability enables the device to improve its energy-saving measures by learning from its decision-making processes and predictive functions.Deep learning, an application based on the artificial neural network method, stands as a powerful tool for the physical layer of 6G. It excels in swiftly resolving computational challenges with access to large information resources. Its rapid iterative operations enable it to tackle maximization problems, minimize computational delays, enhance power efficiency, and offer various other advantages (169).

The synergy between ML (Machine Learning) technology and AI (Artificial Intelligence) will be the key to 6G technology. Building a self aware and adaptive communication network using machine learning is a fundamental aspect of shaping the 6G network. This implementation serves as an adaptable and responsive security layer within the network structure, contributing significantly to its flexibility and dynamism. Furthermore, ML will also enable energy management of mobile networks, leading to notable enhancements in power efficiency, operational efficiency, network configuration, and overall network performance. In essence, ML will be a driving force behind the optimization and functionality of 6G networks.

Research and development in the field of AI is firmly established within 3GPP, and it is anticipated that studies and specifications in Release 18 will drive standards towards incorporating processes that fully leverage the predictive capabilities of data.

The study TR 37.817 on improving data collection for NR (New Radio) and ENDC (E-UTRAN New Radio – Dual Connectivity), with a specific emphasis on three primary AI/ML use cases :

1. Network Energy Saving : This involves activities such as traffic offloading, coverage modification, and cell deactivation to enhance energy efficiency.
2. Load Balancing : The objective is to effectively distribute the network load among cells or specific areas in a multi-frequency/multi-RAT deployment, thereby improving overall network performance through load predictions.
3. Mobility Optimization : This focuses on maintaining satisfactory network performance during mobility events while selecting optimal mobility targets based on predictions of how User Equipments (UEs) may be served.

1.4 Solutions to meet the 5G Expectations and beyoud

To achieve goals such as high data rates, energy consumption reduction, low latency, etc., a combination of various technologies will be necessary. In this context, three paradigms have emerged :

1. **Massive MIMO :** Enhance multiplexing gain and spectral efficiency by implementing an array of antennas at the base station's side. The MIMO technology has attracted a lot of attention in research over the last decade. It was first incorporated into the 3G standard and later into the 4G standard, introducing concepts like multi-user MIMO (MU-MIMO) for single-cell scenarios, and CoMP (Coordinated MultiPoint transmission) for multi-cell scenarios (103). The purpose of this technology is to enhance the benefits of conventional MIMO systems. It refers to a scenario where specific technological components enable the cost-effective deployment of cellular systems utilizing hundreds of antennas at base stations. This approach is used to boost channel capacity and provide substantial gains in multiplexing and diversity for both uplink and downlink directions. These performances will largely depend on the number of antennas at the base station relative to the number of users.
2. **Spectrum sharing :** Transforms the way we allocate and utilize the radio frequency spectrum, aiming to maximize the efficiency of the limited radio frequency spectrum. It facilitates the coexistence of diverse wireless systems and technologies within shared frequency bands, which is particularly fundamental given the escalating demand for high-speed data and seamless connectivity. This approach encompasses a fusion of regulatory guidelines, advanced technologies, and operational strategies meticulously designed to guarantee the equitable and efficient utilization of available spectrum resources. Through dynamic frequency allocation, the implementation of interference mitigation methods, and the coordination of spectrum access, spectrum sharing promotes the overall effectiveness of spectrum utilization, while minimizing the risks of interference. This strategy not only enhances network capacity but also opens doors for novel services and applications, all the while adhering to regulatory guidelines that ensure fair and secure spectrum sharing practices. It helps to increase capacity by utilizing underutilized and unlicensed frequency bands, such as millimeter-wave (mmWave) and unlicensed Long Term Evolution (LTE-U) spectrum.
3. **Ultra Dense Networks :**Increase Spectral Efficiency (SE) by minimizing the distance between the transmitter and receiver while also optimizing frequency reuse. The promising approach to bring this concept into reality is by leveraging Heterogeneous Networks (HetNets) architecture. In HetNets, base stations (BSs) with varying transmission power levels and technologies coexist harmoniously, all with the goal of increasing network capacity. In such a deployment scenario, the traditional macro BS is enhanced by the integration of multiple overlapping tiers of smaller cells, thereby significantly extending the capacity of the system.

   We mainly classify three types of Small Cells based on their transmission power and coverage area :

Femto-cells : These cell types are designed with a maximum emission power of 24dBm to cover a maximum area of 100 meters. Femto-cells are primarily used to ensure coverage for a small number of users within residential environments.

Pico-cells : In this type of cell, the transmission power varies between 24 and 30dBm, and its coverage radius ranges from 200 to 300 meters. Pico-cells can be used to enhance coverage both indoors and outdoors in locations such as hotels, businesses, and more.

Micro-cells : Within this cell category, it's possible to achieve a slightly larger coverage area, extending up to a radius of two kilometers with a maximum transmission power of 40dBm. Like the previous cell types, micro-cells can also be used to ensure indoor and/or outdoor coverage with the potential for up to 2000 simultaneous users.

Table1.1 provides a summary of the various cell types based on their characteristics used in radio communication systems.

TABLEAU 1.1 : Characteristics of the different cell types (10)

Cell type	Power (W)	Coverage radius (km)	Number of users	Area
Femtocell	0,001 à 0,25	0,01 à 0,1	1 à 30	Inside
Picocell	0,25 à 1	0,1 à 0, 2	30 à 100	Inside/Outside
Microcell	1 à 10	0,2 à 2	100 à 2000	Inside/Outside
Macrocell	10 à > 50	8 à 30	>2000	Outside

Standards organizations have reached a consensus that Ultra-Dense Networks (UDN)(97) hold great promise in addressing the ever-growing demand of data traffic. UDN presents the opportunity to provide widespread coverage with strong capacity by making efficient use of available spectrum resources. Additionally, the integration of Small Cells with user devices will significantly reduce latency to less than 1ms and improve energy efficiency due to the shorter wave propagation distances(120). However, the massive deployment of these Small Cells does bring some technical challenges. Firstly, there's the issue of intra-cell interference, and secondly, the proliferation of these small cells could result in increased energy consumption for the network.

Experts have pointed out that such systems tend to be energy-intensive due to their substantial capacity demands. Even though new innovative design approaches like Cloud-RAN (CRAN)(178), Software Defined Networks (SDNs), Network Function Virtualization (NFV)(191), and Mobile Edge

TABLEAU 1.2 : Contributions to Energy Efficiency : A Comparative Overview of CRAN, SDN, NFV, and MEC Technologies

Term	Definition	Energy Efficiency Contribution
Cloud-RAN (CRAN)	The Cloud-RAN (C-RAN) paradigm is recognized for its centralized processing, energy efficiency, real-time computing, and improved spectral utilization. It comprises the Baseband Unit (BBU), Remote Radio Head (RRH), and Optical Transport Network (OTN), enabling base-station functionalities, radiofrequency signaling, and data transmission to the cloud network. Strategically deployed RRHs under the C-RAN architecture enhance scalability and network capacity	CRAN decreases energy consumption by consolidating resources and minimizing the necessity for multiple distributed base station sites, ultimately leading to more centralized and energy-efficient operations.
Software Defined Networks (SDNs)	SDN stands as a pivotal component, facilitating management capabilities for large-scale, high-speed networks by splitting the data plane and control plane. Within the realm of 5G networks, SDN orchestrates and governs applications and services comprehensively across the network, resulting in highly efficient network management.	SDN allows dynamic network administration, enabling the efficient allocation of resources, thereby reducing energy consumption through the minimization of idle resources.
Network Function Virtualization (NFV)	NFV substitutes dedicated network hardware with software-based virtual network functions (VNFs) operating on standardized servers. This transition boosts network flexibility and scalability, resulting in savings of both resources and energy.	NFV enhances energy efficiency by decreasing the reliance on energy-intensive hardware appliances and optimizing resource utilization through dynamic scaling, resulting in more efficient operations.
Mobile Edge Computing(MEC)	MEC also strives to enhance the Radio Access Network (RAN). While CRAN emphasizes centralization and cloud services, MEC takes a different approach, aiming for decentralization by relocating computation, processing, and storage closer to the end user. This approach effectively reduces latency and relieves network congestion in the backhaul network. Initially proposed by ETSI (European Telecommunications Standards Institute) to address network congestion problems, MEC leverages a distributed computing approach.	MEC improves energy efficiency by reducing the need for long-distance data transmission, thereby lowering energy consumption within data centers and network infrastructure.

Computing(MEC)(115) offer more flexibility, control, and network efficiency(as described in table 1.2), there's still a substantial need to reduce their energy consumption. In fact, these design changes aren't primarily focused on making 5G networks more energy efficient. So, counting solely on these approaches and designs to achieve our ambitious goal isn't sustainable and could lead to a serious energy problem.

1.5 Preventing the Energy Crunch

With the upcoming surge in traffic for next-generation cellular systems, there is a pressing need to expand network capacity, requiring the installation of more base stations (BSs). Hence, the proliferation of small cells emerges as a promising solution to fulfill the demands of 5G wireless systems in terms of network capacity and throughput. Integrating small cells of various sizes alongside macro base stations leads to a heterogeneous network (HetNet) that offers high performance and service quality to manage the anticipated rise in traffic. However, the increased number of network components leads to a significant rise in energy consumption. Despite the inherent energy efficiency of small cells, their widespread deployment amplifies both the energy usage of the radio access network and its carbon footprint.

The Information and Communications Technologies (ICT) industry's high energy consumption results in approximately 2% to 3% of global carbon dioxide emissions, accounting for approximately 25% of all emissions (66). Consequently, to significantly reduce current global carbon dioxide emissions in the next-generation wireless networks, the aim is for 5G wireless networks to operate as eco-friendly networks with notably low carbon dioxide emissions. Nevertheless, the designs of conventional cellular wireless networks have primarily prioritized large user throughput and high capacity, with minimal consideration for power or energy efficiency.

Consequently, extensive research is being conducted to minimize BS operations and enhance the network's overall energy efficiency, driven by economic, operational, and environmental considerations (56), (128). As a result, energy efficiency has become a crucial factor for the performance of future 5G networks. With the prospect of millions of additional base stations and billions of connected devices, the importance of energy-efficient system design and operation has grown substantially. It now represents a fundamental aspect in the development of communication networks, marking a shift from emphasizing throughput optimization to prioritizing energy efficiency optimization.

To tackle the imminent energy crisis, prioritizing energy efficiency has become a central focus in the design and operation of future mobile networks. A variety of diverse and complementary approaches have been proposed to improve the energy efficiency of these networks. These methods include optimizing radio resource management, adapting cell configurations, introducing heterogeneous network deployments, and implementing cognitive radio technologies, among others.

1.6 Thesis Outline

The rest of this report is structured as follows. In Section 2, we will provide an overview of specific BS on/off switching strategies in HetNets. Section 3 will present a detailed examination of the power model that forms the foundation for our study. Section 4 offers an extensive exploration of machine learning categories, which underpin our research. In Section 5, we will begin by introducing the proposed metric for optimizing the BS on/off switching strategy, followed by an extensive analysis of our deep learning-based switching control model designed to enhance energy efficiency in 5G networks. Section 6 will describe the simulator designed for our study and analyze the numerical simulation-based system-level results. Finally, in Section 7, we offer concluding remarks and discuss future work directions.

2 Overview of Energy-Efficiency in Cellular Networks

2.1 Introduction

Cellular networks have had a profound impact on our daily lives, and the fifth-generation (5G) of radio technology is set to bring even more transformative changes. It promises to enable unprecedented levels of automation and innovation in various industries due to its enhanced capacity, extensive connectivity, and incredibly reliable low-latency communications. 5G is a complex network that supports a wide range of services, thanks to several key technologies. These include virtualization, which allows for more flexible and efficient network management, softwarization, which makes network functions more adaptable, new Radio Access Networks (RANs) for improved connectivity, and innovative backhaul strategies for better data transport. These technologies work together to provide low latency, high-speed data transfer, and reliable connections. However, as we strive for greater network capacity, broader coverage, and increased data traffic, we're also facing the challenge of rising energy consumption. This isn't sustainable, both from an environmental and a business perspective. Recognizing this, there's a global effort to make our networks more energy-efficient, driven by economic and environmental concerns.

Projections indicate that the number of connected devices could potentially reach 100 billion by 2030(152), and anticipate a significant surge in data traffic, estimated to grow up to 1,000 times more than 4G did in 2018(84). This substantial increase in the usage of smartphones, wearables, and IoT devices is presenting a notable challenge in terms of delivering fast data speeds, extensive coverage, and minimal latency. Furthermore, with each new generation of wireless technology, we have observed a rise in energy consumption due to the addition of hardware to support emerging applications and evolving requirements. The trend suggests that 5G will continue this pattern, significantly increasing energy consumption compared to its predecessor, 4G. The need to accommodate high data rates and a multitude of connected devices is making these networks more energy-intensive. In fact, it is estimated that 5G's energy consumption could be four times that of 4G(77). Consequently, the pursuit of energy efficiency emerges as a critical concern in the context of 5G, distinguishing it from earlier generations. To illustrate the scale and significance of the 5G enabling impact, it's worth highlighting that the International Telecommunication Union (ITU) SMART 2020 report (89) estimated that mobile communications alone contributed approximately 2,135 million tons of CO2e emissions in 2018. Furthermore, predictions from (87) indicate that this impact is expected to grow substantially during the 5G era. By the end of 2020, it was projected that the cumulative enabling effect across the entire Information and Communication Technology (ICT) sector would amount to approximately 15% of the world's total greenhouse gas emissions.

In light of this issue, the Third Generation Partnership Project (3GPP) has introduced the New Radio (NR) specification, which equips the next generation of networks with tools to significantly reduce energy usage and greenhouse gas emissions. This aligns with the broader sustainability goals of the Information and Communication Technology (ICT) sector, contributing to a more environmentally responsible and efficient future.

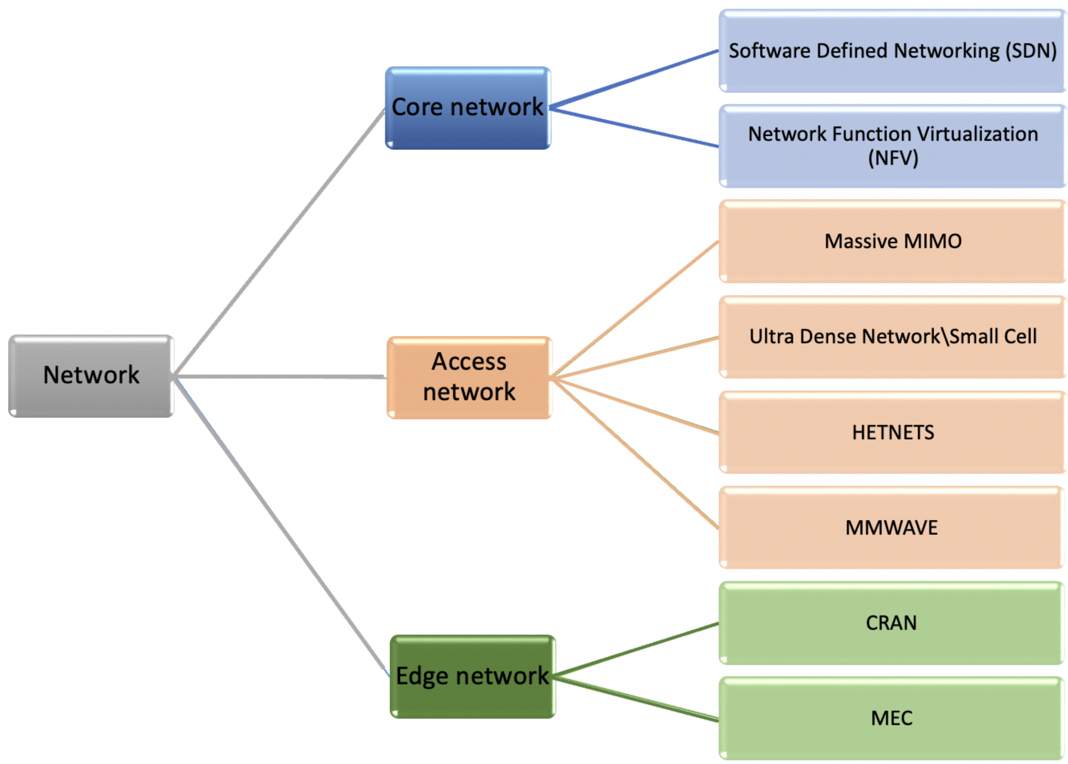


FIGURE 2.1 : 5G technologies designed to boost energy efficiency.

As well, in the development of 5G networks, several advanced technologies are being integrated to support a wide range of services. These technologies encompass Software-Defined Networking (SDN), Ultra-Dense Networks (UDN), Network Function Virtualization (NFV), mobile edge computing (MEC), and cloud computing. However, this integration of diverse technologies introduces challenges related to energy efficiency. For instance, in the case of Ultra-Dense Networks (UDN), while energy consumption benefits from reduced transmission power in less dense scenarios, the rise in computational demands leads to higher energy usage in denser environments (Choras). This computational intensity is expected to grow steadily over time. Additionally, to meet the increasing demand for connectivity, massive MIMO (Multiple Input, Multiple Output) technology is employed to serve densely populated areas. However, achieving the right balance between linearity and efficiency is crucial in massive MIMO systems. The performance of power amplifiers has a direct impact on the energy efficiency of these systems. Striving for linearity in power amplifiers can lead to increased costs, while embracing non-linearity may have a detrimental impact on energy efficiency. Given the high demands of future technologies, addressing these challenges requires the development of suitable hardware, the implementation of intelligent energy-efficient decision-making techniques, and the formulation of innovative network designs to effectively manage energy consumption.

2.2 Techniques for Enhancing Energy Efficiency in 5G Networks

2.2.1 Algorithms and Strategies for Energy-aware Cellular Networks

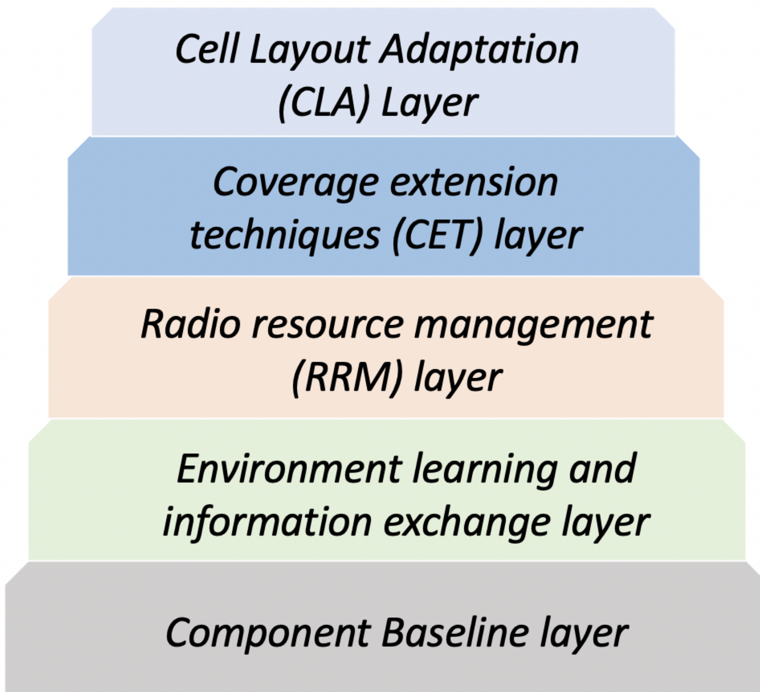


FIGURE 2.2 : Framework with Stacked Layers for Energy-Efficient Strategies in Cellular Networks

As mobile cellular networks' energy demands continue to rise, many researchers have directed their efforts toward enhancing the energy efficiency of these systems. A significant early contribution to understanding how existing cellular networks could operate with energy conservation in mind can be found in the work of (40). In this study, the authors introduced a framework for assessing the potential to save energy and maximize energy efficiency in three key areas of mobile radio networks : components, links, and network levels. They emphasized that achieving higher energy efficiency doesn't rely on optimizing just one aspect or component but on optimizing all these areas together. To improve our insight into how these diverse elements of mobile networks are linked to the goal of reducing energy consumption, (155) introduced a model for energy-efficient techniques. This model is structured like a stack of layers, where improvements in the lower layers result in more significant savings in the upper layers (as shown in Figure 2.2). These approaches can be categorized into the following groups :

• **Component baseline layer :** This layer serves as the foundation for improving RAN (Radio Access Network) energy efficiency. Tasks carried out at this layer encompass adaptive power amplification, energy-efficient hardware, and the use of adaptive radiation patterns through beamforming. Enhancements at this level result in reduced energy consumption of the Base Station's (BS) components, providing flexibility in design constraints and facilitating the operations of higher layers. Nevertheless, focusing solely on component-level improvements is insufficient for achieving significant energy savings due to resource under-utilization, highlighting the need to address the upper layers.

• **Environment learning and information exchange layer :** In the subsequent layer, Cognitive Radio (CR) is introduced to enhance spectral efficiency (SE) by adjusting the transmission parameters of radio devices according to external environmental conditions. The concept of Cognitive Radio aims to intelligently identify unused spectrum bands and swiftly adapt to them by reconfiguring transmission and reception parameters in accordance with the channel's characteristics.

• **Radio resource management (RRM) layer :** This layer encompasses mechanisms like power control, Discontinuous Transmission (DTX), antenna adaptation, and radio resource allocation. Techniques in this layer focus on regulating transmission power and optimizing the allocation of transmission resources across time and bandwidth. Within this layer, researchers explore trade-offs between spectrum/energy, bandwidth/power, and delay/power.

• **Coverage extension layer :** This layer extends to HetNets (Heterogeneous Networks) and relay systems. In such networks, Radio Resource Management (RRM) proves highly beneficial as it facilitates collaboration between various network tiers, leading to additional energy and cost savings.

• **CLA layer :** Lastly, at the highest level of the hierarchy, the Cell-Layout Adaptation (CLA) encompasses cell size adjustments, including techniques like cell breathing and switching-off. It's worth noting that this layer has the potential to yield greater energy savings compared to the lower layers, as we will explore further.

Building upon the energy efficiency framework outlined above, there are various approaches and techniques available for enhancing the energy efficiency of forthcoming 5G cellular networks. Previous research (6) has categorized these approaches into five broad categories as follows :

1. Optimizing the energy efficiency of the radio transmission process
2. Hardware-based approach
3. Base station policies
4. Embracing the use of harvesting energy sources.
5. Network planning and deployment of heterogeneous networks.

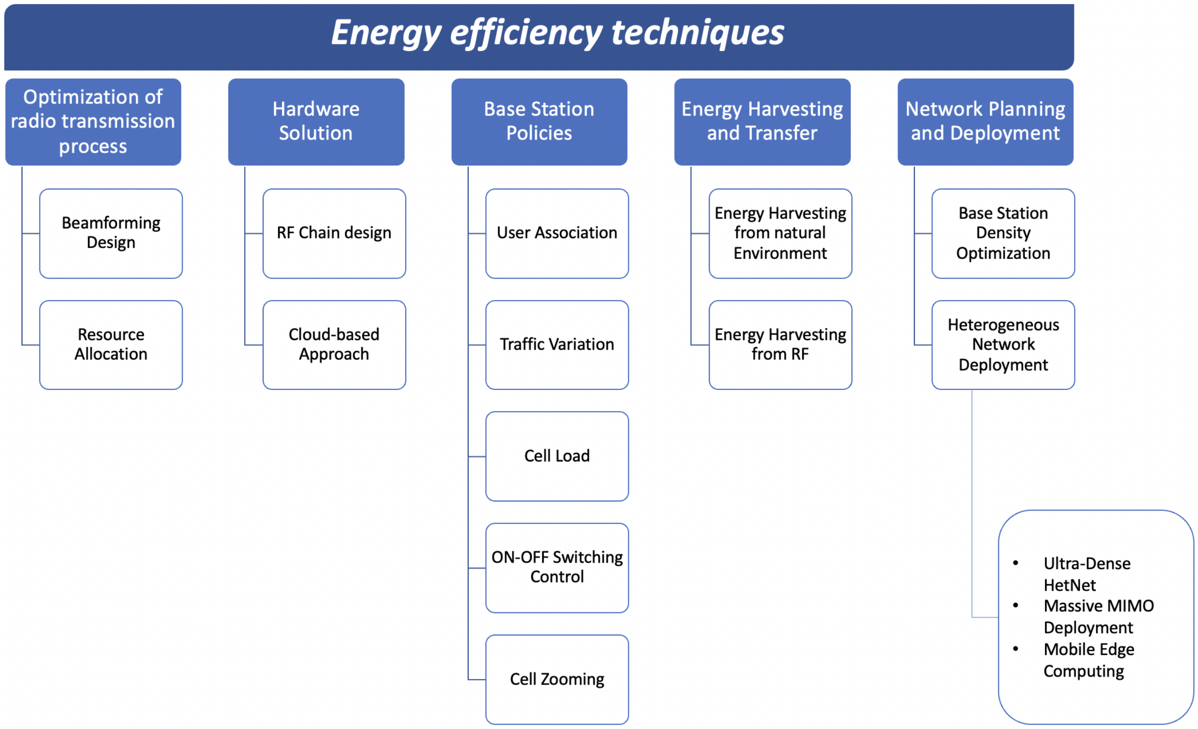


FIGURE 2.3 : Strategies for Energy Optimization in Wireless Mobile Networks

2.2.1.1 Optimizing the Energy Efficiency of the Radio Transmission Process

The first category prioritizes the optimization of the radio transmission process, encompassing methods that target aspects at either the physical or MAC (Media Access Control) layer. Various novel technologies, including cognitive radio transmission, Massive MIMO,channel coding, and resource allocation , are being explored to optimize energy consumption in telecommunication networks(75). The strategies, explored in the following, primarily focus on enhancing signal transmission patterns, network coverage, and interference reduction through the optimization of radio transmission parameters.

Beamforming : In the realm of MIMO HetNets, beamforming is an effective tool for improving network efficiency by reducing interference (116). Interference can severely impact network energy efficiency (EE)(142), making beamforming a key strategy to enhance EE. Beamforming relies on RF precoding units, focusing on signal processing methods, including linear techniques like minimum mean square error (MMSE) , zeroforcing (ZF), and maximum ratio combining or transmission (MRC/MRTT), and non-linear methods like vector perturbation, and dirty paper coding (DPC) (28). Due to their simple implementation, linear methods are often preferred. Research in (168) explores the use of DPC for energy-efficient designs and ZF, proposing centralized and decentralized cooperative beamforming for multi-user MIMO and mMIMO systems. However, these studies lack comparative analysis with other beamforming approaches. This limitation is addressed in (24), which evaluates the energy effeciency performance of ZF,MMSE, and MRC in mMIMO systems, demonstrating ZF's superiority due to their ability to reduce complexity while handling interference suppression. While MRC is computationally efficient, it performs poorly in EE due to its inability to handle interference and noise. In contrast, MMSE effectively deals with both but suffers from high signaling overhead. These insights underscore the widespread adoption of ZF in energy-efficient beamforming designs.

Effective resource allocation : is crucial for wireless network performance, particularly in terms of energy efficiency (EE) (187). Unfair allocation can negatively impact users with weaker channel conditions, diminishing EE. In scenarios with overlay spectrum access, insufficient bandwidth allocation to small base stations (SBSs) can worsen the problem (161), with overlay and underlay spectrum access concepts. Addressing this issue, "boundary allocation" groups users with similar channel conditions to identify and compensate the worst-performing group. Furthermore, An optimization problem(206) is presented to maximize energy efficiency within the absolute blank subframe (ABS) scheme, with defined upper and lower bounds., ensuring a minimum quality of service (QoS) for the worst-served users.

An alternative method for improving EE resource allocation in HetNets involves the implementation of soft frequency reuse(SFR), which reduces inter-cell interference at cell edges to improve the user's signal-to-interference-plus-noise ratio (SINR). SFR divides each cell into cell center and cell edge regions, with exclusive bandwidth portions and varying transmit power levels to meet data rate requirements. It guarantees orthogonal sub-bands for the cell edge region to avoid interference. SFR's energy-saving potential lies in BSs reducing transmission power through optimal power allocation and interference reduction at cell edges (207). The investigation of energy-efficient network design through the implementation of SFR is addressed in (85). An issue with SFR is limited spectrum reuse, leading to the development of "multi-layer frequency reuse" (MSFR), which divides cells into more than two regions. This improves SINR for users in intermediate regions while maintaining favorable conditions for edge users (69).

This study emphasizes transmission power allocation, a key factor in base station (BS) power consumption. Optimizing transmission power assigns more power to channels with high gain and less to channels with poor reception, resulting in varying quality of service (QoS). Several works on power allocation strategies exist under orthogonal multiple access (OMA)(161)(21) and non-orthogonal multiple access (NOMA)(52)(62). NOMA gains attention for serving multiple users simultaneously, and becomes notably pertinent in the context of network design prioritizing energy efficiency. It multiplexes users based on different power levels, allowing users to be multiplexed by power levels and mitigating inter-user interference using successive interference cancellation (SIC).

Various approaches have been proposed to effectively utilize resources in time, frequency, and spatial domains to achieve energy savings. Importantly, these approaches are cost-effective and often do not necessitate hardware replacements. However, it's worth noting that trade-offs between energy efficiency and other network performance metrics are likely to arise. Additionally, challenges related to measuring errors, stemming from complex uncertainties like noise and interference, have yet to be fully addressed.

2.2.1.2 Hardware-based Approach

The second category focuses on enhancing hardware components, such as power amplifiers, with greater energy efficiency, as seen in references (38). Many components used in today's cellular network architecture are suboptimal in terms of energy efficiency. To reduce the power consumption of both base stations (BSs) and user equipment (UEs), it's essential to simplify their hardware designs. (86) outlines the complete analysis of 5G application-oriented hardware design for UE. However, the primary concern in wireless mobile networks is conserving energy at the BSs.

Research in this field has been directed for designing RF hardware and developing networks using cloud technology. Within RF hardware design, diverse arrangements of RF chains and antenna units have been suggested, particularly in the context of mMIMO utilization. Another critical aspect is the power efficiency of components like power amplifiers (PAs). For instance, in a typical cellular BS, where more than 80% of input energy dissipates as heat, the PA stands out as the largest energy consumer. Generally, the useful output power accounts for only about 5% to 20% of the input power. Studies suggest that the power efficiency of PAs, represented by the ratio of output power to input power, could potentially reach as high as 70% (176). This implies substantial energy savings if more energy-efficient components are adopted in the network.

However, it's essential to consider the implementation cost of these approaches, which can be quite high. For example, a power amplifier module with 35% power efficiency, designed for small cell WCDMA or LTE BSs (covering a radius of up to 2 km), can cost around $75. The cost would be even greater for larger coverage areas or higher power efficiency requirements. Consequently, network operators must carefully weigh operational and economic factors before deciding on hardware replacements. The cloud-enabled strategy, commonly referred to as Cloud Radio Access Network (CRAN), offers a way to simplify the hardware of base stations (BSs) by offloading many computational tasks from the BS to a centralized remote location(140). In CRAN, The processing of the baseband signal and the management of the RF unit are divided into two distinct components : the Baseband Unit (BBU) and the Remote Radio Head (RRH)(111). The Baseband Units (BBUs) are integrated within the cloud network and linked to the Base Stations (BSs) via backhaul connections in distant coverage regions. Regarding Heterogeneous Networks (HetNets), the RAN (radio access Network) that has evolved from this methodology is recognized as H-CRAN (Heterogeneous Cloud Radio Access Networks).

Efforts to enhance the energy efficiency (EE) of CRAN are evident in studies like(8). In the context of CRAN's downlink, (41)investigate energy efficiency by comparing strategies involving data sharing and data compression for transmission. Data sharing involves distributing the message for a specific user across multiple base stations (BSs) in the network, employing collaborative beamforming techniques to transmit it to the targeted user. These methodologies aim to minimize transmission power while meeting user target rates. The results demonstrate that the data sharing strategy exhibits superior performance at lower user target rates, mainly because of the decreased power consumption in the backhaul. However, at higher data rates, the condensed method is favored as it lowers backhaul power usage.

The suggestion put forward by (8) is to incorporate software-defined networking within CRAN, referred to as SD-CRAN. The power model is constructed through the assessment of the energy consumption of each specific SD-CRAN component. The findings reveal that SD-CRAN consumes more power than traditional CRAN. Nevertheless, the network scalability advantages of SD-CRAN make it a fitting choice for deployment in ultra-dense networks..

With the growth of networks, the computational intricacy of both SD-CRAN and CRAN is predicted to rise. Moreover, the challenge lies in addressing throughput latency, primarily caused by the distance between BSs and the core network. Mobile Edge Computing (MEC) technology has been effective in mitigating latency issues. However, Given the possible benefits and limitations of these cloud technologies concerning network scalability, the power used for computation, and the delay in data service, it is crucial to examine these elements before determining the most suitable deployment, whether it involves SD-CRAN, CRAN, or MEC.

2.2.1.3 Base Station Policies

The third category involves strategies focused on Base Stations (BSs) during periods of low traffic, as BSs are the most significant energy consumers in cellular networks. A commonly employed method for enhancing coverage and capacity in mobile networks is the implementation of a hierarchical cell structure (HCS). This HCS employs a structure comprising Macrocells, Microcells, and Picocells, as depicted in Figure 2.4. While the HCS strategy effectively extends network coverage and capacity, it can also lead to increased energy consumption. The fundamental cell structure is designed to accommodate peak traffic loads, often resulting in over-dimensioning during non-peak hours.

Traffic Variation : From an energy optimization perspective, a well-explored approach involves deactivating or putting cells to sleep when their capacity is underutilized. (26) employs user equipment (UE) traffic patterns to generate traffic forecasts within their respective cell areas. This paper demonstrates that energy savings can be achieved through forecasting methods, and these savings are closely tied to the density of base stations. However, these methods rely solely on traffic forecasts when reactivating base stations, which is a static and inefficient approach. This can lead to network congestion, and both solutions are highly sensitive to changes in coverage area, making scalability and real-world implementation challenging due to varying attenuation conditions.

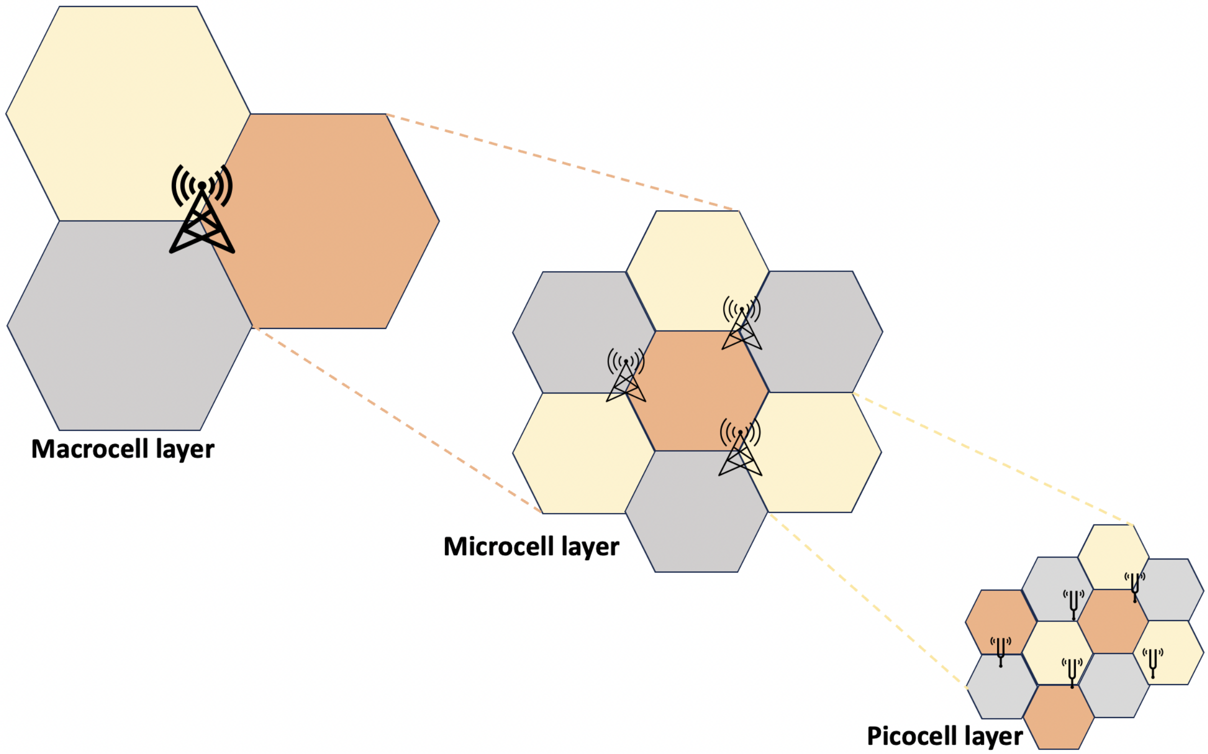


FIGURE 2.4 : HEC structure

In contrast, (113) introduces a performance metric known as the area spectral efficiency (ASE) to assess area utilization. They utilize the Signal-to-Interference-plus-Noise Ratio (SINR) from connected UEs to estimate ASE, and a simplified version based on Signal-to-Noise Ratio (SNR) measurements is also proposed. When the number of active UEs falls below a certain threshold, base stations are deactivated. Conversely, when the ASE in the area decreases, indicating more active UEs, base stations are reactivated. Simulations show that this approach can reduce the energy consumption of their test network by 60-70%. However, this solution lacks the capability for UE handovers in cells scheduled for deactivation, potentially affecting Quality of Experience (QoE). Moreover, there is no clear indication of the solution's scalability.

Modeling the traffic patterns of User Equipment (UE) allows for the prediction of optimal periods to deactivate Base Stations (BSs) during specific hours. However, relying solely on periodic on/off triggers may not be ideal, especially during peak periods when BS outages can occur suddenly. Therefore, integrating a dynamic traffic strategy that factors in different UE activities, including UE velocity, becomes increasingly feasible for network planning. Dynamic traffic models have been applied in energy-efficient design scenarios, as showcased in (70**?**), where queuing theoretic models are employed, (133), which utilizes Markov decision processes, and (51), which is formed on the basis of genetic algorithms' principles.

Cell zooming : within extremely dense HetNets, employing dynamic traffic profiles for BS deactivation might elevate energy usage in the network because of heightened interference levels caused by the excessive allocation from nearby base stations.(109).However, with the cell zooming technique, a base station (BS) in HetNet can dynamically modify its transmission power and, thus, alter its cell size in response to changes in traffic(203).To achieve efficient UE association, the transmission power can be amplified based on the design goal, enabling the shutdown of lightly utilized cells while upholding coverage. (184). Conversely, to mitigate interference levels within the network, transmission power can be reduced. An alternative scenario involves reducing transmission power aligned with how UEs move to minimize energy usage(150).Figure 2.5 provides illustrations of cell zooming through coverage reduction and cell sleep-mode.

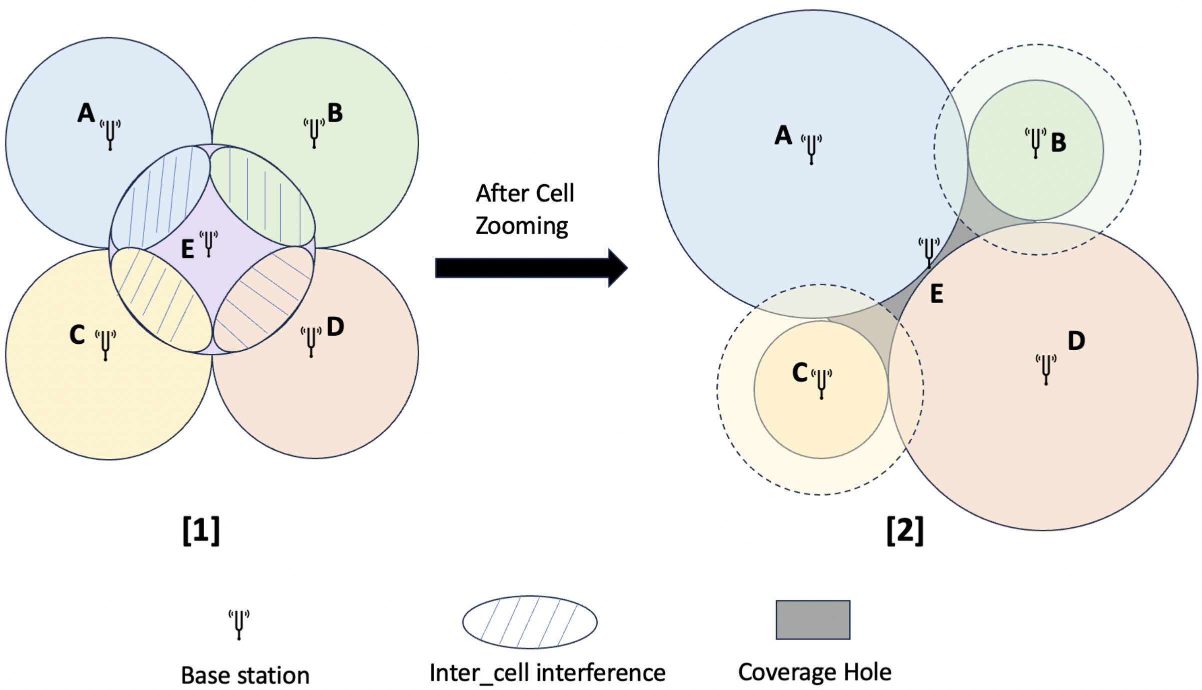


FIGURE 2.5 : Cell zooming [1] : Normal cells deployed A-E. [2] : Cell E in sleep mode, cells A and C increasing their coverage to compensate, and B and D reducing their coverage.

The cell zooming algorithm, as introduced in (127), operates through a defined zooming period encompassing three stages. First, there's a coordination phase in which base stations within a specific area gather cell-related data. Next, a transition phase follows, during which base stations make adjustments based on the coordination findings. Finally, a serving phase ensues, where base stations function within the new configuration. A comparison was made between a centralized and a distributed version of this algorithm, revealing that the centralized approach outperforms the distributed one in terms of energy consumption relative to outage blocking probability. In the centralized model, a cell zooming server is employed for collecting and coordinating base station settings, although this may raise scalability concerns.

The challenges linked to cell zooming are explained in (127). Foremost among these challenges is the impact of user equipment (UE) movements, which can induce traffic fluctuations leading to compensation loops. A second challenge lies in compatibility, as implementing cell zooming necessitates a substantial base station upgrade to support functions like tilt adjustment and new management channels. The last challenge pertains to unintended network behavior, exemplified by inter-cell interference (as depicted in Figure 2.5, [1]) or gaps in coverage (as exemplified in Figure 2.5, [2]).

Some of these challenges are addressed in (173), where it is proposed that the introduction of more base stations can enhance energy savings. In this scenario, additional base stations can be put in sleep mode during off-peak hours, but the viability of this approach heavily hinges on the specific traffic patterns within the area. It's worth noting that the energy consumption increases with the number of active base stations, and the cell zooming approach primarily relies on transmit power adjustment as its sole method, assuming a uniform distribution of traffic and interference, which may not hold true in real-world implementations.

User Association : Efficiently associating user equipment (UE) with specific base stations (BSs) plays a crucial role in optimizing the BS sleep mode technique within a Heterogeneous Network (HetNet). Effective UE-BS association allows for the identification and deactivation of lightly loaded BSs. UE will link to the BS with the highest RSRP (Reference Signal Received Power). One of the commonly used techniques to achieve efficient user equipment (UE) connection in HetNet is the cell range expansion (CRE) feature of SBSs, referred to as biasing (190).

Moreover, investigations into CRE within ultra-dense HetNets have revealed its potential to reduce unwarranted handovers for users on the move, thereby enhancing the UE connection procedure(163). Nonetheless, UEs prioritize connections to Macro Base Stations (MBS) due to their greater transmission power in the absence of coordination in the CRE process. To tackle this issue, various solutions have been proposed. These include allocating a sub-frame of Absolute Blank Subframe (ABS) scheme for Small Base Station (SBS) operation (162) and restricting the transmission power of MBS on a sub-frame to reduce interference (45).

Recent contributions in this area have taken into account additional factors such as the random distribution of BSs and UEs (202), Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) propagation effects (187), traffic awareness (106), UE mobility and handovers (63), data transfer delay (46), and equitable resource distribution(102).

Cell Load Considerations : Before deactivating a base station (BS), it's crucial to assess the load status of neighboring cells. Failing to do so may lead to a degradation in Quality of Service (QoS) for user equipment (UE) transferred from a sleeping cell to active neighboring cells, potentially resulting in dropped connections. Relevant research on this matter is outlined in (32).

In (16), an algorithm is introduced to identify candidate cells for deactivation, employing a model based on packet traffic for estimating the load. Following the identification phase, the algorithm estimates the effect on neighboring cells and selects the BS with the least impact as the preferred candidate for deactivation. In contrast, the algorithm presented in (166)is designed to deactivate small cells (SBS) experiencing the highest interference levels while avoiding the deactivation of cells serving the highest number of users. Additionally, (32) contributes to this area by employing a discrete-time Markov process to create an effective traffic diversion plan aimed at minimizing energy usage.

Base Station (BS) ON/OFF Strategies : Strategies for turning base stations ON and OFF can be categorized into two main types : partial ON/OFF, where a fraction of radio resources is activated or deactivated, and complete ON/OFF, where the BS is either fully operational or completely powered down (210). However, practical implementation of ON/OFF switching strategies requires careful consideration of their impact on service latency and energy usage. The effect of service delay on user equipment (UE) associated with BSs in sleep mode, along with the associated energy efficiency trade-offs, is explored in (126). These studies introduce various Wake-up techniques utilizing the vacation queue model.

Recent research has delved into the effects of switching on energy consumption, with notable studies including those by [(83)–(61)]. (83), who propose an enhanced Markov process for traffic prediction to reduce frequent switching in areas with fluctuating traffic patterns. (42) introduce a central controller based on fuzzy logic to coordinate small cell base stations (SBSs), preventing simultaneous activation of SBSs with DTX (discontinuous transmission) features and mitigating sudden increases in interference levels. In contrast, (14) present a distributed scheme utilizing a non-cooperative regret-based satisfaction game technique, eliminating the need for a central controller. Moreover, (61) investigate the collective impacts of initial energy consumption costs arising from frequent SBS switching, formulating an optimization problem for energy reduction that considers factors such as BS density, UE mobility patterns, and the transmission power of UEs.

We will talk and explore in more detail in the following section about on-off strategies for base stations, since this category has the potential to achieve greater energy savings compared to other categories.

2.2.1.4 Embracing the Use of Harvesting Energy Sources.

The pursuit of reducing CO2 emissions and minimizing the need for frequent UE recharging has given rise to an innovative energy-efficient technique known as energy harvesting and transfer. This represents the fourth category of energy-efficient approaches, which harness renewable energy sources. In contrast to conventional energy sources like hydrocarbons, which contribute to greenhouse gas emissions, renewable resources such as hydro, wind, and solar power are notable for their sustainability and environmentally friendly nature (35). Telecom manufacturers have devised plans to supply cellular base stations (BSs) powered by solar energy in underdeveloped regions. These areas often lack well-maintained roads and reliable access to traditional energy sources like diesel, making off-grid BSs a challenge to power.

Energy harvesting techniques involve harnessing the available energy from renewable resources to supplement existing electric-powered infrastructure. This approach holds promise as a long-term environmental solution for the mobile cellular network industry, particularly in areas without well-established network infrastructure. In contrast, for developed countries with mature infrastructure, similar concerns arise regarding the embodied and replacement costs, much like the component-based approaches. Transitioning services from outdated electric-powered BSs to new energy harvesting BSs presents technical challenges in maintaining fault tolerance and data security without service interruptions. Research on this energy-efficient technique can be categorized into two groups : energy harvesting from the natural environment and energy harvesting from the radio environment.

Energy Harvesting from the Natural Environment : Recent studies have emphasized the economic and ecological performance of renewable energy sources in wireless cellular networks(67), (170). However, sustainability can be a challenge due to varying climate conditions. Integration with conventional grid energy sources using smart grid technology has been proposed to address this. Analyzing CO2 emission reduction from the perspective of minimizing energy procurement costs, researchers have investigated the performance of genetic algorithms (GA) and particle swarm optimization (PSO)(67) in deactivating redundant base stations. Results suggest that while the PSO approach outperformed GA, purchasing energy from renewable sources reduced profits and increased CO2 levels(170).

The concept of base station cooperation(181)has also been introduced to minimize energy procurement costs, with schemes such as energy, communication, and hybrid cooperation. The hybrid scheme displayed superior performance in energy cost savings. Strategies integrating radio optimization techniques, such as transmission power and sub-carrier allocation optimization (170), and joint coordinated multipoint techniques (90)for interference avoidance, have also been explored to reduce energy consumption rates from renewable sources.

Energy Harvesting from Radio Frequency : Efforts in wireless cellular networks are directed towards achieving high data rates (197) while mitigating the increased power consumption of mobile devices. The concept of Simultaneous Wireless Information and Power Transfer (SWIPT) is proposed (138)to serve as a continuous energy source for mobile terminals. Tractable models are developed to evaluate performance in both the uplink and downlink of SWIPT HetNets, accounting for energy harvested(5).

Studies propose mixed beamforming techniques (151) to enhance energy harvesting, yet the integration of information decoding and energy harvesting within a single device remains a challenge (160). A power-splitting Zero Forcing (ZF) receiver is introduced (53) to concurrently receive information and harvest energy, eliminating the need for separate devices. However, the study does not address the system's EH and ID efficiency.

In the design of SWIPT networks(93), the power consumption model and computational complexity of solution algorithms are crucial. A distributed beamforming approach is introduced to reduce signaling overhead, demonstrating convergence with minimal iterations. The use of linear and non-linear power consumption models is also explored, indicating limitations and non-linear characteristics of power amplifiers at high frequencies.

2.2.1.5 Network Planning and Deployment of Heterogeneous Networks

The last category addresses the issue by incorporating small cells into the cellular network, including micro, pico, and femto cells. These compact cells are designed to serve small areas with dense traffic, employing low-power cellular base stations (49) that are both cost-effective and often support plug-and-play functionality. In contrast to the conventional deployment of homogeneous macro cells, this heterogeneous deployment strategy reduces energy consumption in the network by reducing the propagation distance between network nodes and utilizing higher frequency bands to support increased data rates.

However, there are significant challenges associated with these approaches. The introduction of additional small cells can result in increased radio interference compared to conventional homogeneous macro cell networks, potentially affecting the quality of the user experience negatively. Furthermore, if an excessive number of micro, pico, or femto cells are deployed, the energy-saving trend may be reversed due to the additional embodied energy consumed by the newly deployed cells and the overhead introduced in transmission. Therefore, careful planning is required to determine the optimal number and locations of these smaller cells in order to achieve a reduction in total energy consumption. Furthermore, the combination of heterogeneous network deployment with sleep mode strategies has demonstrated the potential to yield significant energy savings, as indicated by previous studies in (154).

The efficient planning and deployment of cellular networks can be examined by considering various aspects, such as base station deployment strategies and established techniques commonly used in Heterogeneous Networks (HetNets). These approaches can be categorized and discussed as follows :

Optimizing base station (BS) density : is a crucial aspect of planning mobile cellular networks, involving the delineation of the network coverage area of interest (164). Once the coverage area is defined, estimating the number of BSs needed to ensure a specific level of Quality of Service (QoS) becomes possible (13). In dense Heterogeneous Networks (HetNets), where small cell BS (SBS) deployment may be unplanned and irregular (71), devising simplified and practical frameworks for optimal BS density planning is essential. Various solution methods have been explored to address this challenge, including metaheuristic approaches (68), greedy algorithms (154), and stochastic geometry (209), which have proven to be valuable.

In the study presented in (68), a metaheuristic approach involving particle swarm optimization (PSO) and grey wolf optimization (GWO) methods is used. These methods integrate data rate and coverage constraints into the algorithms to optimize the global population and deactivate redundant base stations (BSs). However, it may not be suitable for Ultra-Dense HetNets (UD-HetNets) due to its uniform distribution of small base stations (SBSs).

The greedy-based algorithm proposed in (154) gradually adds SBSs until the average spectral efficiency reaches a defined condition, deactivating redundant BSs during off-peak traffic periods. While it extends the BS distribution topology to a random one suitable for UD-HetNets, it lacks explicit analytical foundations for the chosen topology.

To address these limitations and provide practical analyses of ad hoc BS planning and deployment, stochastic geometry is utilized. Zhou et al. (209) employ a Poisson distribution process to model User Equipment (UE) distribution and generate diverse traffic patterns. A heuristic algorithm then evaluates the state of BSs and associated UEs, estimating the optimal number of BSs to deploy. However, the use of a centralized processing scheme (154) may impact scalability due to increased signaling overhead.

In another related design by Alkan (44), SBS deployment planning is based on the reported signal strength of Main UE (MUE) pilot channels. However, this approach does not thoroughly address pilot channel information processing, which is critical due to potential challenges with false information arising from pilot contamination.

In summary, for an effective BS optimization approach in planning energy-efficient UD-HetNets, certain critical considerations must be taken into account. These include ensuring a minimum required QoS, employing suitable analytical models for various topologies, accounting for traffic dynamics, and implementing efficient algorithms.

Heterogeneous Network (HetNet) Deployment : In HetNets, the reduced transmission distance between Small Base Stations (SBSs) and User Equipment (UEs) contributes to a notable enhancement in network Energy Efficiency (EE). This improvement is further supported by the inherent characteristics of millimeter Waves (mmWaves), facilitating spatial densification of SBSs and spectrum densification via massive Multiple Input, Multiple Output (mMIMO) technology (110). Moreover, the adoption of edge computing, which involves offloading signal processing and content delivery from the cloud to edge devices, minimizes the separation between the core network and edge devices, leading to significant energy savings (182). These strategies will be discussed in detail in the subsequent sections.

a) Ultra-Dense (UD) Base Stations (BSs) : Studies on Ultra-Dense Heterogeneous Networks (UD-HetNets) suggest that Energy Efficiency (EE) and Spectrum Efficiency (SE) are expected to increase up to a certain threshold. However, further increasing the Base Station (BS) density beyond this threshold can lead to the degradation of both EE and SE (185). This raises questions regarding the optimal level of network densification to enhance EE without compromising SE, and how to jointly optimize EE and SE. Analytical tools such as the Poisson Point Process (PPP) have been utilized to establish parameters for developing frameworks that characterize network EE and SE (104), focusing on coverage and association probabilities.

While some studies, such as (201), have focused on estimating traffic loads for each network tier, they primarily provide a closed-form expression for a global EE model without comprehensive insights into SE. Other works, including (104), examine both EE and SE, emphasizing the impact of SBS density on EE and measuring SE against Signal-to-Interference-and-Noise Ratio (SINR), although this approach doesn't fully elucidate the effect of SBS density on SE. Research such as (33) proposes joint optimization of SE and EE but doesn't evaluate them based on SBS density.

To address these limitations, recent studies, including (113),(185), employ heuristic procedures and cooperative game approaches to jointly optimize area SE and EE, revealing that beyond a certain SBS density, EE starts to decline due to cumulative circuit power consumption, while SE continues to rise until a point when interference causes it to decrease. The Poisson Cluster Process (PCP) is suggested as a more suitable approach than PPP for modeling BS distributions in UD-HetNets (2). However, the application of PCP in energy-efficient network design requires further exploration. Recent research, as in (96), introduces a variant of PCP, the Matern cluster process, to formulate expressions for spectrum efficiency and power consumption minimization problems, highlighting the impact of the received interference by UEs in a cluster.

Overall, these studies emphasize the intricate relationship between EE and SE in HetNets, underscoring the need to control SE within acceptable limits when pursuing EE objectives in dense networks.

b) Massive MIMO Deployment : Although Energy Efficiency (EE) achieved through BS densification encounters limitations due to interference from nearby BSs, the potential for exceeding this limit is evident through massive Multiple-Input, Multiple-Output (mMIMO) deployment (25). By leveraging mMIMO, power levels for both BSs and User Equipment (UEs) can be reduced without compromising Quality of Service (QoS) (1). This makes mMIMO a promising solution for enhancing both EE and Spectrum Efficiency (SE) in wireless communication systems.

Practical design considerations for energy-efficient mMIMO systems encompass power consumption in the Radio Frequency (RF) unit and computational power, both of which scale with the increasing number of antennas (65). To address these challenges, hybrid precoders have emerged as an alternative to digital and analog precoders, aiming to reduce energy consumption in the RF unit, address hardware complexities, and improve beamforming accuracy (159).

Among hybrid precoders, partially-connected structures are favored for energy-efficient designs, with low-complexity algorithms like successive interference cancellation (SIC) and partial singular value decomposition (SVD) using Givens transformations providing optimal precoding vectors (64), (112). While these schemes enhance EE, the SE of fully-connected hybrid precoders remains higher due to increased beamforming gain (112).

Novel solutions like replacing phase shifters in the analog unit with phase over-samplers (POS) and switches have demonstrated improved EE performance in hybrid precoders (108). Additionally, adaptive operations of hybrid precoders through optimal configuration of precoding components have shown potential for further enhancing EE, especially in high Signal-to-Noise Ratio (SNR) regions (208). Overall, these studies underscore the critical role of hardware complexity and configuration in determining the EE and SE of mMIMO systems.

c) Mobile Edge Computing (MEC) : Mobile Edge Computing (MEC) technology brings cloud computing closer to User Equipment (UEs) and Base Stations (BSs), reducing latency and computational complexity (4; 158). While conventional MEC focuses on computation offloading, recent studies highlight the benefits of caching in enhancing Energy Efficiency (EE) (107). Various strategies such as multi-user computation offloading games and energy-efficient offloading techniques emphasize the preference for local computation under high interference, resulting in lower energy consumption and shorter processing times (31; 201). Addressing the quantification of EE, (156) formulates a joint optimization problem for energy-efficient computation offloading, showing the effectiveness of local computation for small data sizes and offloading for larger data sizes. Moreover, investigations into the impact of caching on energy-efficient MEC architecture (94; 78) reveal the potential of joint cooperative and coded caching to minimize energy consumption and improve the performance of virtual reality applications. These findings emphasize the significance of caching for optimizing energy-efficient MEC systems.

2.2.2 Review of HetNet's BS On/Off Switching Approaches

Reducing power consumption in cellular heterogeneous networks by dynamically turning off base stations has recently gained increasing attention. The literature has introduced several methods for designing energy-efficient networks in the context of wireless mobile operations. Many of these methods concentrate on enhancing the energy efficiency (EE) of base stations (BS), as BS units are responsible for a substantial portion, ranging from 60% to 80% (90), of the total energy consumption in wireless mobile networks. Consequently, several studies have been established to improve the energy efficiency of the network. As such, BS on/off switching algorithms are among the most powerful energy-saving solutions. These algorithms switch off unnecessary parts of the network (i.e., BS) and offload users to neighboring BSs during off-peak periods. They also activate the appropriate number of BSs as required. Therefore, BS on/off switching algorithms would be particularly advantageous for future fifth generation (5G) networks characterized by extreme BSs densification, making the energy efficiency challenge even more complex.

Base Station Switching (ON/OFF), which represents one of the potential strategies for enhancing network efficiency, holds significant promise but poses notable implementation challenges. The complexity arises from the fact that deactivating base stations results in a complete cessation of services within a given area, particularly when considering a single-tier coverage model. However, this obstacle can be overcome through the introduction of 5G networks, which are anticipated to expand the radio interface infrastructure in various sizes and hierarchies, forming heterogeneous networks (HetNets). This HetNet paradigm is expected to create overlapping networks comprising larger Macro Base Stations (MBSs) covering smaller underlying Small Base Stations (SBSs).

Organizing heterogeneous networks in this manner levels the playing field for base station switching algorithms by increasing the overall network power requirements, making the technology essential, while simultaneously facilitating its implementation through the overlapping coverage. Before this paradigm shift, for a base station to be powered off, it had to possess a mechanism to monitor service requests effectively to prevent coverage gaps. In contrast, within HetNets, deactivating an idle SBS primarily results in a loss of potential capacity, which, when well-planned, can contribute to energy savings without disrupting the network's regular operations. As HetNets become more widely adopted, the persistent challenge of avoiding complete service outages in an area is effectively resolved, allowing for the assumption that BS switching is feasible only in regions with overlapping coverage.

Several BS switch-on/off strategies (76) have been proposed from different design perspectives to only optimize energy savings or other energy- related performance trade-offs, such as random, load-aware, and distance-aware, strategies. Besides, some research has been undertaken to consider a common design of the BS switch-off strategy and other strategies, such as user association, physical-layer interference cancellation strategies, and resource allocation. BS switch-off strategies could be achieved by focusing either on small base stations (SBS) or macro base stations (MBS) or on both types of BS in HetNets. However, deactivation of MBS which is derived based on minimizing BS power consumption can have a significant negative impact on network coverage. In contrast, the random deactivation strategy of SBS is designed based on the maximization of energy efficiency (EE) with the constraint for coverage probability. A lot of work has been done to design a specific BS switch-off strategy in HetNets. Nevertheless, improvements and challenges remain to be explored by appropriately exploiting the combination of different criteria to get greener and better network performance. There are generally two approaches to properly switch on/off a given number of BSs (174),(144) : offline and online approaches. The relatively simple offline approach allows the preplanning in advance of the on/off switching intervals. However, it has a major drawback, namely that it does not take into account the current instantaneous load and, hence, is not robust to unpredictable events (i.e., failure, random hotspot, etc.), which therefore restricting its efficiency. On the other hand, the online approach exclusively considers the actual load (i.e., instantaneous) to decide whether to switch off a BS or not. Thus, it can handle unpredictable events by activating the appropriate part of the network (i.e., the number of BSs) to cope with an unexpected traffic increase or decrease. Despite its importance, this approach is complex since the decision must be made and executed in real-time. Besides that, it may incur a large number of on/off switching operations with important energy costs which may not only considerably increase the overall network energy cost, but also damage hardware equipment.

Certain studies concentrate on the utilization of mathematical optimization (57) to manage the activation and deactivation of Base Stations (BS) in a manner that maintains the user experience while simultaneously decreasing energy usage. These algorithms switch off unnecessary parts of the network (i.e., BS) and offload users to neighboring BSs during off-peak periods. They also activate the appropriate number of BSs as required (36). Therefore, BS on/off switching algorithms would be particularly advantageous for future fifth generation (5G) networks characterized by extreme BSs densification, making the energy efficiency challenge even more complex.

To this end, investigations have focused on energy savings when the network is lightly loaded, and the stations can be put on sleep mode. Most of these structures exploit the evolution of traffic during the day : when the load is low for a sufficient period (especially at night), some stations are switched off and their users are taken care of by those who remain in service (27),(100). Different techniques are implemented in this direction to ensure that the service is nevertheless satisfactory. These approaches demonstrate significant advantages in terms of both energy consumption and bit rate. However, it's important to consider that a cell is a dynamic environment where users constantly change their positions and numbers. Therefore, time holds paramount importance, yet these methodologies consume excessive time and computational resources.

Nowadays, the advancements in Artificial Intelligence (AI) have led to widespread adoption of machine learning techniques for optimizing wireless communications. These techniques have been integrated into self-organizing networks (SON), With the objective of making daily tasks for network operators more streamlined and efficient (141). By incorporating AI-based tools, the progression of the SON paradigm in 5G shifts towards a proactive methodology. This methodology capitalizes on the vast amount of available data and incorporates additional dimensions derived from characterizing end-user experience and behavior (60).

In this regard, various energy-efficient schemes utilizing machine learning have been developed. These approaches demonstrate significant advantages in terms of both energy consumption and bit rate (58). The main strategies for optimizing energy efficiency at the base station level include adaptive sleep modes based on reinforcement learning (50), traffic-based switch-off strategies (73), and efficient transmission power control(180). However, it's important to consider that a cell is a dynamic environment where users constantly change their positions and numbers. In addition, these ML-based methods do not yield energy savings to the same degree as the mathematical approach, and they also involve a substantial cost.

Optimizing base station switching still presents certain challenges that require further attention. In response to these challenges, there is a wealth of scholarly activity aimed at exploring various possible implementations. Many of these implementations identify various factors contributing to complexity. However, the primary issue with switching base stations lies in the fact that current equipment is not designed to frequently accommodate changes in operational mode and demands special consideration during the switching process. This limitation discourages service providers from employing the approach frequently. The inability to switch off BSs on-demand necessitates precise estimation of future service requirements for a given area, a complex problem that relies on a deep understanding of the area's service history and the use of sophisticated estimation tools. Additionally, the challenge of solving the combinatorial optimization problem, which determines the most efficient combination of BSs to activate at a given time, becomes more complicated as the number of combinations grows.

Given these scenarios, where optimization methods require substantial time and computational resources, and Machine Learning (ML) approaches haven't achieved a satisfactory balance between cost and savings, we recognized the potential of combining the strengths of both fields. Our objective is to achieve significant enhancements in energy efficiency while simultaneously decreasing the time and computation demands. This motivated us to develop a deep learning approach centered around solving optimization problems in wireless networks taking into account different criteria and perspectives. However, an adaptable design that can respond effectively to changes in network conditions would result in greater energy savings and improved network performance.

2.2.3 When EE meets AI

Beyond transmitting data with enhanced energy efficiency, the next generation of wireless networks faces the formidable challenge of accommodating a wide array of use cases, each with distinct service level requirements, all within an exceedingly dynamic environment. Meeting these demands is no small feat, given the increasing complexity, heterogeneity, and constant evolution of the network landscape. This necessitates the development of innovative smart wireless radio technologies, sophisticated spectrum management techniques, and adaptive decision-making mechanisms to effectively cater to these diverse requirements.

Furthermore, the sheer volume of reconfigurable parameters within future networks is staggering. To illustrate, consider the progression from 2G to 4G networks, where the number of such parameters surged from 500 in 2G to 1000 in 3G and 1500 in 4G, as reported in (88). This figure is projected to soar even higher in 5G networks, exceeding 2000. Consequently, enhancing the intelligence of these networks becomes imperative to realize the Self-Organized Network (SON) paradigm, characterized by self-configuration, self-optimization, and self-healing capabilities.

Figure 2.6 shows the development of an intelligent radio system with the ability to learn from past experiences or observations. This system aims to improve its overall performance by establishing a utility function for each action taken, relying on many reputable sources such as (95), (134) and (198). To realize this vision, it is imperative to seamlessly integrate fundamental concepts from the fields of artificial intelligence (AI) and machine learning (ML) into wireless infrastructure and end-user devices.

AI represents a computational framework that imparts intelligence to machines, allowing them to learn, operate, and respond in a manner akin to human behavior. The roots of this paradigm trace

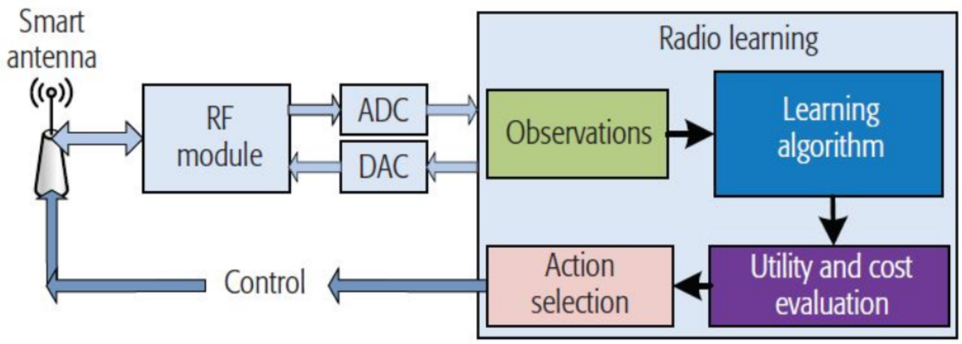


FIGURE 2.6 : The concept of intelligent radio learning(95)

back 75 years, when threshold logic was employed to formulate a computational model for neural networks, as highlighted in (139). However, it wasn't until the late 1980s that neural networks began to pique interest, albeit with gradual progress due to computational power limitations. Today, thanks to significant advancements in Graphics Processing Units (GPUs), AI has garnered unprecedented attention and is making inroads into a multitude of computer science domains.

Within the expansive AI landscape, Machine Learning (ML) has emerged as a particularly promising technique for absorbing knowledge from data and making informed decisions, predictions, and recommendations. This is achieved without the need for explicit programming, as underscored by (95) and (198).

Artificial intelligence (AI) has evolved to encompass various discipline techniques, including machine learning, optimization theory, game theory, and meta-heuristics [168]. Among these, machine learning has emerged as a pivotal subfield of AI. In the context of future 5G networks, machine learning holds significant potential for orchestrating and managing network resources. The integration of intelligence can offer more efficient solutions to technical challenges in next-generation systems, such as device-to-device (D2D) communication, large-scale massive MIMO, and the management of heterogeneous networks with diverse technologies and architectures.

Dynamic radio cell operation stands as a pivotal technology in the pursuit of Energy Efficiency (EE). An essential goal within this domain is the optimization of radio resource allocation through actions like cell activation/deactivation or cell zooming, all of which respond to the ever-changing traffic demands. The key to making this optimization a reality lies in acquiring an in-depth understanding of dynamic user traffic patterns and mobility behaviors over time. Moreover, it entails the ability to predict future traffic demands and mobility patterns by learning from historical data. Mobile network functions inherently hold a wealth of user context information, encompassing mobility histories, state transitions, traffic patterns, and more. This wealth of data serves as a valuable resource for uncovering underlying rules and optimizing the EE of mobile networks.

The actual environment-related information concerning energy availability, coupled with practical user behaviors, is far from purely random. By leveraging diverse sets of information encompassing energy availability processes, user mobility, link quality, traffic characteristics, and Quality of Service (QoS) requirements, we have the potential to proactively tailor network resource allocation to meet user demands effectively.

2.3 Conclusion

In reviewing the pursuit of energy-efficient wireless mobile HetNets, various angles have been explored, ranging from optimizing Base Station (BS) deployment and improving radio transmissions to managing BS sleep patterns and utilizing renewable energy sources. However, these efforts have not been without challenges. Designing energy-efficient HetNets involves addressing various critical factors that present our research challenges. These encompass managing randomly distributed Small Base Stations (SBS), making optimal decisions on UE (User Equipment) associations, dynamically allocating resources adapted to the fluctuation of traffic patterns, and respecting the trade-off between energy efficiency and the high-quality service (Qos) standards.

3 Machine Learning Overview

The journey of crafting intelligent programs, also known as artificial intelligence (AI), commenced in the 1950s. Machine learning, a subset of AI that doesn't require explicit rule-based programming to acquire knowledge, began its evolution in the mid-1980s and steadily progressed over the years. While developers initially focused on crafting algorithms to solve specific problems, the emphasis gradually shifted towards the creation of learning algorithms capable of autonomously solving unsolved problems through data-driven methods. These algorithms are commonly known as Machine Learning (ML) algorithms. ML algorithms typically aim to optimize the performance metrics of a parametric model on given datasets to acquire the ability to tackle complex problems. It's essential to highlight that ML algorithms fundamentally differ from the functional models they seek to parameterize effectively. Among the various AI models, Artificial Neural Networks (ANNs), often referred to simply as NNs, have gained substantial popularity. These mathematical frameworks draw loose inspiration from biological neurons. NNs are constructed by combining numerous local elementary operations, known as neurons, to represent highly complex global functions. These models, sometimes called connectionist models, offer efficient hardware implementations and are exceptionally well-suited for applying ML techniques. This versatility has led to their successful deployment across a wide array of domains and a growing interest from various research fields, including their potential applications in future generations of cellular networks. This hierarchy of concepts is illustrated in Figure 3.1.

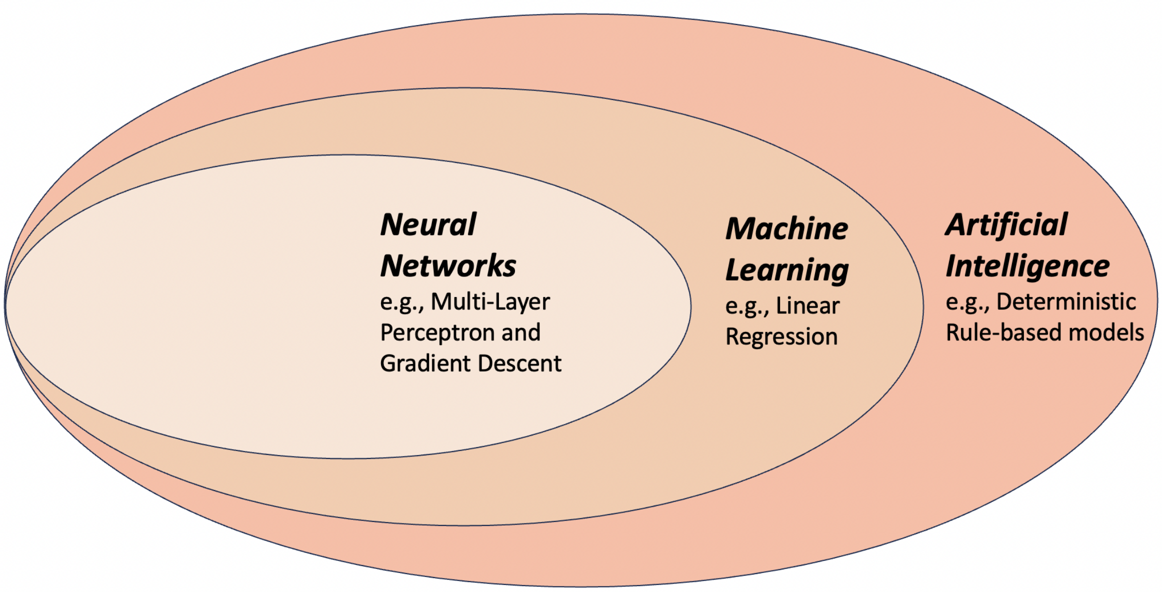


FIGURE 3.1 : A brief hierarchy of artificial intelligence concepts.

Machine learning, also known as automated learning, is a subfield of artificial intelligence that has brought about a profound transformation in our ability to extract insights and make informed decisions from data. At its core, this discipline relies on computers' ability to learn from experience, discern patterns, and progressively enhance their performance without the need for explicit, handcrafted instructions. In essence, machine learning empowers machines to adapt and evolve based on the data they encounter.

The significance and practical applications of machine learning span a wide array of domains. Whether it is deployed in medicine, finance, industry, logistics, advertising, or even our day-to-day interactions with technology, machine learning assumes a pivotal role. It facilitates more informed decision-making, automates complex tasks, optimizes processes, and fosters the development of smarter, more personalized products and services. In summary, machine learning holds immense promise for enhancing efficiency, accuracy, and innovation across numerous sectors.

Machine learning has evolved significantly from its fundamentals to become a sophisticated discipline capable of handling complex tasks. This development can be attributed to rapid advances in computing power, which allows the training of more complex models, and to the wealth of available data, which serves as raw material for these algorithms. Additionally, machine learning has undergone a transition from traditional statistical methods to deep learning, characterized by neural networks with many layers that excel at processing unstructured data such as text, images, and data. audio. As the field continues to advance, ethical considerations and responsible use of AI are gaining importance, paving the way for a new era of discussions regarding fairness, transparency, and accountability in learning applications automatic.

Accordingly, machine learning algorithms have gained significant attention recently for addressing various challenges across multiple domains, including resource management, power allocation, cell sleeping, and precoding. In this section, we will explore the diverse range of machine learning algorithms employed to enhance energy efficiency in wireless networks. Additionally, we will provide a concise overview of the benefits of utilizing machine learning approaches over traditional methods to enhance energy efficiency in 5G and future-generation networks."

3.1 Basics of Machine Learning

3.1.1 Categories of Learning

Machine learning is categorized into three primary branches : supervised learning(SL), unsupervised learning(USL), and reinforcement learning(RL). Deep Learning, a subset of machine learning that emerged around 2010, also encompasses these three categories supervised, unsupervised, and reinforcement learning. Machine learning classification techniques and learning algorithms (as indicated in the figure 3.2) are a prevalent choice in the context of both 5G enabling technologies and addressing some concerns like energy efficiency.

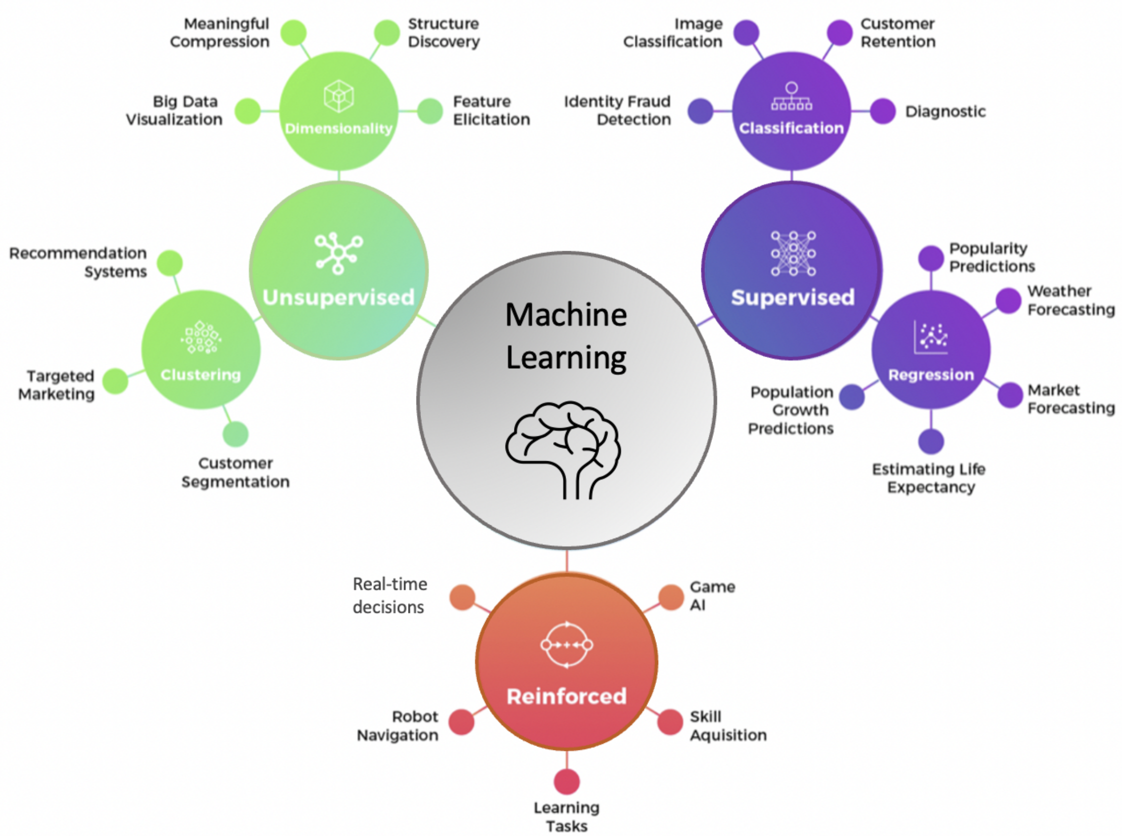


FIGURE 3.2 : Types of ML algorithms.

Supervised learning involves providing the agent with a dataset consisting of labeled input examples and their corresponding desired outputs, which are supplied by an external supervisor with knowledge in the field. Each example in this dataset comprises a specific situation and the correct action associated with it. The primary goal of supervised learning is to establish the connection between variables in order to deduce a general rule that can map inputs to outputs through extrapolation and regression. Supervised learning is particularly well-suited for channel-related problems such as channel estimation, detection, and learning behavior for future predictions. This is because supervised learning leverages past data to generate outputs, relying on previous experiences as a foundation.

Unsupervised learning,in contrast to supervised learning, focuses on identifying concealed structures within sets of unlabeled data. In this category, the agent is tasked with recognizing patterns in its input data without the guidance of labeled examples. Unsupervised learning plays a significant role in cellular networks, where it is commonly employed. Well-known unsupervised techniques like Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are utilized to unveil correlations and hidden structures among variables. For instance, the PCA method is employed to reduce the complexity of massive MIMO systems by decreasing their reception matrices. Furthermore, clustering techniques, which fall under the category of unsupervised methods, demonstrate their usefulness in the detection of anomalies within the network. Also, unsupervised learning differs from supervised learning in its suitability for wireless network clustering and spectrum sensing challenges. It autonomously learns the network and tackles complex issues independently, making it a valuable choice for solving intricate problems beyond the scope of supervised learning.

Reinforcement learning, differs significantly in its approach to training an agent. Instead of relying on pre-annotated datasets, reinforcement learning operates through interaction with a dynamic environment. Without any prior knowledge of system information, the reinforcement learning process aims to maximize a numerical reward signal by making decisions at each time step. This involves mapping states to actions that yield the highest possible reward. Reinforcement learning finds extensive application in decision-making scenarios, including radio resource management and user selection within cellular networks. Within this category, Q-learning stands out as a commonly used technique to address such problems. Reinforcement learning, on the other hand, excels in scenarios where the problems are not well-defined, such as resource allocation and management in networks. It has the capability to adapt its approach to achieve desired outcomes, systematically learning from results and refining its decisions.

3.1.2 Types of Data used in Machine Learning

In the domain of machine learning, the careful selection and meticulous preparation of data stand as pivotal factors in determining the success of predictive models and analytical tasks. These data types primarily fall into two categories : structured and unstructured data. A profound comprehension of these data types and their distinctive characteristics holds immense importance when it comes to making informed decisions while designing machine learning algorithms and models.

3.1.2.1 Structured Data

Structured data encompasses information that follows a systematic and predictable order. This category of data is commonly presented within tabular or relational databases, where each data element resides in a well-defined field or column. The inherent organization of structured data lends itself effectively to traditional statistical analysis and the utilization of machine learning algorithms.Illustrative instances of structured data encompass :

Numerical Data : This category includes integers and floating-point numbers, rendering it suitable for mathematical operations. Instances include stock prices, temperature measurements, or individuals' ages...

Categorical Data : Categorical data represents distinct categories and is frequently employed for object or concept classification. Models include product categories (e.g., electronics, apparel) or customer segments (e.g., platinum, gold, silver).

Date and Time Data : Date and time values serve the purpose of tracking events and trends across time, making them crucial for time series analysis. Illustrative examples include timestamps of online purchases, sensor data, or social media posts.

Structured data's well-organized nature makes it highly compatible with traditional machine learning algorithms, such as linear regression and decision trees.

3.1.2.2 Unstructured Data

Conversely, unstructured data refers to information that does not have a distinct, pre-defined structure. Its distinctive features lie in its versatility and unpredictability with regard to both form and content. Unstructured data can appear in various forms, including text, images, audio, video, and other formats. It presents considerable challenges for conventional data analysis methods due to its complexity, but is gaining growing significance in the realm of machine learning. Illustrations of unstructured data comprise :

Text Data : This category includes documents, articles, emails, and social media posts. To extract valuable insights from text data, Natural Language Processing (NLP) techniques come into play, enabling applications like sentiment analysis, chatbots, and language translation.

Visual Data : Images and videos are repositories of rich information but necessitate the application of computer vision techniques for interpretation. These techniques find applications ranging from facial recognition to the analysis of medical images.

Audio Data : Audio files, consisting of sound waves, serve as valuable sources for speech recognition, voice assistants, and music recommendation systems.

Sensor Data : In the realm of IoT (Internet of Things) applications, unstructured sensor data is prevalent and can encompass information from various sensors, including temperature sensors, GPS, and accelerometers.

Leveraging unstructured data often entails the utilization of deep learning methods, such as convolutional neural networks (CNNs) for images and recurrent neural networks (RNNs) for text and speech analysis.

To conclude, structured and unstructured data serve as the foundational pillars of machine learning applications. Gaining a comprehensive grasp of their distinct characteristics and disparities is imperative when it comes to selecting suitable preprocessing methods, feature engineering approaches, and machine learning algorithms. Proficiency in managing both structured and unstructured data is indispensable for researchers and practitioners, enabling them to effectively address a broad spectrum of real-world challenges. In practice, many datasets consist of a combination of both types, necessitating a diverse tools and approaches to extract valuable insights and patterns.

3.1.3 The Main Stages of the Machine Learning Process

The machine learning workflow generally comprises several key phases, which may exhibit minor variations depending on the particular problem and methodology. As stated by Ashmore et al. in their work (17), the process of creating a machine learning-based solution in an industrial context can be grouped into four primary phases (132) :

Data Management : This stage revolves around the preparation of the data required for constructing a machine learning model. Data is a fundamental part of every machine learning solution. The success of the solution relies on both the algorithm and the quality of the training and testing data. Therefore, creating high-quality datasets usually marks the initial phase of any operational machine learning pipeline. There are three key steps to data management : data collection, data preprocessing, and data exploration.

— **Data Collection :** Data collection is the first step. It's about finding and understanding the data available and figuring out where to store it conveniently. The data can be collected from various sources like databases, files, APIs, or sensors.

— **Data Preprocessing :** encompasses several essential tasks aimed at refining and structuring raw data for machine learning. These activities involve identifying a data schema, handling missing values through imputation, simplifying and ordering the data, and converting it from its raw form into a more convenient format. Additionally, data preprocessing may entail encoding categorical variables and other essential operations to ensure the data is prepared adequately for machine learning tasks.

— **Data Exploration :** often referred to as EDA or Exploratory Data Analysis, entails the practice of visually examining and analyzing data to acquire insights and enhance comprehension of its inherent attributes. During this phase, patterns, correlations, outliers, and other significant information within the data can be identified, leading to a deeper understanding of the dataset.

Model Learning : Here, the selection and training of the machine learning model occur.

— **Model Selection** in machine learning constitutes a pivotal undertaking that revolves around the identification of the most appropriate algorithm or model structure for a given task. This process involves a nuanced equilibrium between the intricacy of the model and its performance characteristics. the process of model selection typically involves a profound understanding of the problem domain and the available dataset. Researchers and practitioners must meticulously consider a multitude of factors, encompassing the dataset's scale, the dimensionality of the feature space, computational resources. A diverse array of algorithms, ranging from straightforward linear regression to intricate deep neural networks, are at one's disposal, each bearing its own strengths and limitations. Ultimately, the model choice must harmonize with the precise objectives and constraints inherent to the machine learning project, ensuring the delivery of dependable and meaningful outcomes.

— **Training :** During the training phase, a machine learning algorithm or model is employed to discern patterns and connections within the data. This model is exposed to a subset of the dataset known as the training set, enabling it to acquire the ability to make predictions or classifications based on input features.

Model Evaluation : The primary objective of this phase is to ensure that the model meets specific functional and performance criteria.

— **Evaluation :** Following the training phase, it becomes imperative to evaluate the model's performance. Typically, this is achieved by utilizing a distinct section of the dataset (known as the validation or test set) that the model has not been exposed to during training. The selection of evaluation metrics depends on the nature of the problem and commonly includes accuracy, precision, recall, among others.

— **Refinement :** Subsequent to the evaluation findings, there may arise a need to add Hyperparameters that include parameters such as (the depth of a decision tree, the number of hidden layers within a neural network, or the quantity of neighbors in a k-Nearest Neighbors classifier,...) or to make adjustments to enhance its performance. This is an iterative process that could involve experimenting with diverse algorithms, conducting feature engineering, or implementing regularization techniques.

Model Deployment : This stage involves integrating the trained model into the necessary software infrastructure to facilitate its execution. Additionally, it encompasses considerations regarding model maintenance and updates.

— **Deployment :** After achieving a level of satisfaction with the model's performance, it can be rolled out to a production environment, allowing it to provide predictions for new, unobserved data. The rollout process may encompass the integration of the model into various platforms, including web applications, mobile apps, or other software systems.

— **Monitoring and Maintenance :** Following deployment, it is crucial to continuously monitor the model's performance in real-world scenarios. This involves tasks like evaluating its accuracy, identifying instances of concept drift (shifts in data distribution), and making necessary model updates to uphold its continued effectiveness.
— **Updating :** Over time, as fresh data becomes available, it is often essential to retrain the model to keep it current and preserve its accuracy. This can be carried out at regular intervals to incorporate the most recent information.

In summary, the machine learning process functions as a repetitive cycle, dependent on continuous feedback and refinement. This recurring methodology is vital for enhancing the model's performance and ensuring its alignment with evolving data and shifting requirements.

3.2 Artificial Neural Networks Strategies for Energy Efficiency

The emerging wireless technology paradigm demands high data rates and supports a wide array of applications, challenging traditional technology in terms of learning and decision-making processes. Here are some of the advantages of machine learning over conventional approaches :

— Machine learning significantly enhances learning speed, especially for large-scale problems, as it can adapt and learn from its data, whereas older techniques are typically hardcoded.
— Machine learning has the ability to acquire knowledge from data, while traditional techniques are predominantly based on fixed, predefined rules.
— Machine learning possesses autonomous decision-making capabilities, whereas traditional systems require new sets of instructions for each new function.
— The development of software for new applications is often a costly and time-consuming endeavor.

However, alongside these advantages, there are also disadvantages to consider when it comes to machine learning, particularly in the context of training, large-scale processing, security, and the implementation of research theories at the application level.

Due to the unpredictability in mobile environments regarding information accuracy and the full understanding of how the environment will evolve, researchers have turned to the model of reinforcement learning framework. This approach is employed to tackle stochastic optimization challenges within wireless networks. Specifically, in scenarios involving sequential decision-making, reinforcement learning is employed to discover the best strategy by actively engaging with the uncertain environment through interaction.

Within wireless cellular networks, traditional decision-making problems encompass resource allocation, user scheduling, and sleep schemes. These tasks carry substantial computational complexity and are of utmost significance due to the dynamic variations in the environment, such as shifting traffic patterns, interference levels, energy availability, and electricity prices. Finding ways to efficiently utilize resources represents a newly emerging challenge. In addition to conventional methodologies and tools, considerable attention has recently been directed towards Q-learning solutions (136)(3). For instance, distributed Q-learning-based algorithms have been proposed in the context of HetNets, where cells learn when to enter a sleep state through interactions with the environment. The primary objective of these agents, represented by BSs, is to minimize energy consumption while simultaneously upholding system performance. In studies like (118),(183), the network is conceptualized as a Multi-Agent Reinforcement Learning (MARL) system, where each small cell employs a distributed Q-learning algorithm to determine an optimal policy independently, without reliance on other cells. Nevertheless, due to the absence of coordination among the base stations (BSs) employing distributed Q-learning, further work was conducted as an extension in (136). In addition to reinforcement learning, the authors introduced an additional centralized layer based on neural networks. This layer is responsible for augmenting the distributed Q-learning within each small cell by providing supplementary information that influences the local actions to be taken. These algorithms belong to a category known as Heuristically-Accelerated Multiagent Reinforcement Learning (HAMRL)(119).

Conventional reinforcement learning encounters a scalability challenge. In simple models characterized by a limited number of states and actions, reinforcement learning techniques demonstrate efficiency in discovering optimal policies. However, when dealing with more intricate environments and tasks, classical reinforcement learning methods fall short. Recently, Deep Q-Learning (DQL), leveraging deep neural networks, has proven to be a successful approach in augmenting the learning capability of reinforcement learning for complex tasks (189).

In the research outlined in (122), the authors illustrated the effectiveness of Deep Q-Learning (DQL) in the context of dynamic power allocation within wireless networks. Their primary objective was to optimize the weighted-sum rate utility function. To achieve this, they introduced a distributed dynamic power allocation scheme based on DQL, designed to be scalable for extensive networks. This approach stands out from existing solutions that typically deal with complex optimization problems. Instead, the authors emphasized how DQL has the potential to effectively address large-scale network challenges that conventional optimization tools struggle to tackle.

3.2.1 AI/ML Algorithms Designed for Load Prediction

Given the diverse set of features involved in loading prediction, such as current and historical load data, as well as neighboring cells' load information, various techniques are being explored to tailor the prediction model to the specific characteristics of each feature. The research in the literature focused on predicting network traffic can be categorized into two distinct groups based on the methods employed : statistical-based approaches and machine learning-based approaches.

3.2.1.1 Statistical-based Approaches

Statistical-based approaches involve the analysis of network traffic statistics. One commonly used statistical method for predicting network traffic is the Autoregressive Integrated Moving Average (ARIMA) model(20). ARIMA combines the autoregressive and moving average models to make predictions by considering past time-series values while accounting for non-stationarity. However, ARIMA has limitations, particularly in capturing the seasonality (repeating cycles) of network traffic. To address this limitation, an extension called Seasonal Autoregressive Integrated Moving Average (SARIMA) has been introduced(186). Despite their utility, statistical methods like ARIMA and SARIMA struggle with capturing rapid traffic fluctuations since they primarily rely on historical data mean values. Furthermore, these methods tend to be linear and may not provide high accuracy when dealing with the complex and dynamic traffic patterns often observed in real network scenarios.

3.2.1.2 Machine Learning-based Approaches

ML-based techniques have emerged as an alternative to statistical methods for network traffic prediction. These data-driven approaches have garnered attention due to their ability to effectively model non-linear relationships and leverage the wealth of data collected by base stations (BSs).

However, traditional ML algorithms like k-nearest neighbors (KNN)(117) and support vector regression (SVR)(175)present certain challenges. They necessitate meticulous parameter tuning to attain precise predictions. Moreover, these methods are characterized by a limited memory span, attributed to their constrained parameter sets and computationally intensive nature, which can limit their potential to enhance prediction precision.

Subsequent research has shifted towards the utilization of recurrent neural networks (RNNs) to effectively model complex nonlinear sequential patterns. This approach has yielded promising outcomes across diverse domains, including speech recognition, image captioning, and natural language processing. Notably, the introduction of the Long Short-Term Memory (LSTM) cell has addressed the issue of vanishing gradients encountered in traditional RNNs (192). A variant of LSTM, known as Convolutional LSTM (ConvLSTM), was introduced to enhances spatio-temporal data processing by incorporating convolutional neural networks (CNNs)(199), replacing dense connections. Effective for traffic prediction, it treats data as images but relies on grid-based partitions, limiting applicability.

Long Short-Term Memory (LSTM) : is a type of artificial recurrent neural network (RNN) used in deep learning. Unlike standard convolutional neural networks, LSTM can process both individual data points and sequences of data, making it suitable for tasks such as video or speech analysis. With its multiple memory cells, LSTM can retain information over lengthy sequences, making it ideal for time series data, natural language processing, and speech recognition. Its strength lies in preserving context over extended periods, making it well-suited for various complex tasks.

The LSTM structure facilitates the understanding of long-term relationships within input time series data. A distinctive feature of an LSTM unit is its three gates : the input gate, forget gate, and output gate. These gates control the unit's functions, considering inputs such as the current input vector, memory from the previous time-step, and the output from the previous time-step. Non-linearity is managed through a blend of sigmoid and hyperbolic tangent units, implementing their respective functions. For instance, in Figure 3.3, a neural network structure consisting of stacked LSTM units is illustrated.

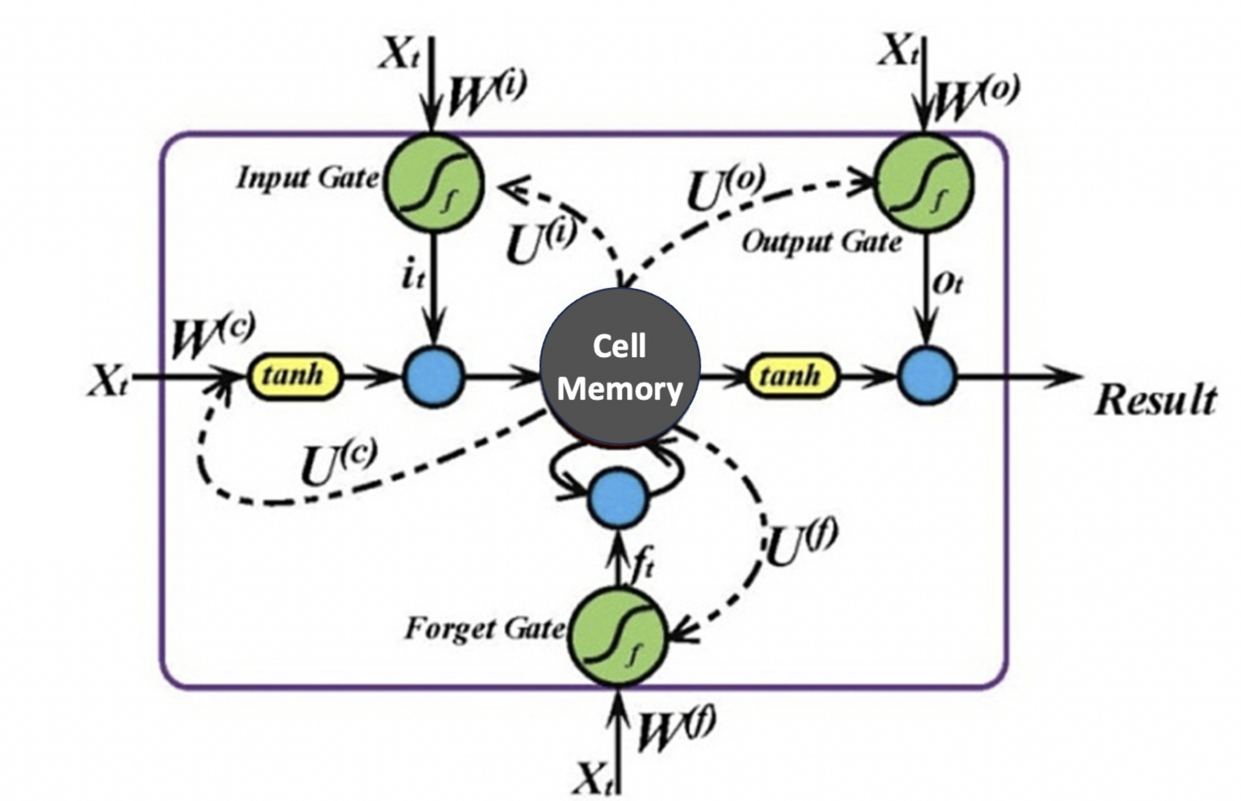


FIGURE 3.3 : The structure of LSTM's memory unit.

The Long Short-Term Memory (LSTM) neural network comprises several essential elements :

1. **Input Layer :** This initial layer receives sequential data as input, which can include time series data or textual information.
2. **LSTM Cells :** The LSTM network consists of multiple LSTM cells, each featuring three crucial gates :

   Input Gate : Responsible for regulating the flow of new information into the cell.

Forget Gate : Manages the decision to retain or discard information from the cell's memory.

Output Gate : Dictates which information should be propagated to the subsequent time step or output.

3. **Cell State :** This component serves as the memory of the LSTM cell and extends across time, enabling the network to selectively retain or discard information as required.
4. **Hidden State :** The LSTM cell's output, often used for making predictions or relayed to subsequent LSTM cells in a recurrent manner.
5. **Output Layer :** In various applications, an output layer is introduced to facilitate predictions or data classification, leveraging the insights gleaned from the LSTM cells.
6. **Activation Functions :** Sigmoid and hyperbolic tangent (tanh) activation functions are frequently employed within LSTM cells to govern the behavior of the gates and cell state.

At the core of the unit lies a memory cell, represented by the gray circle, while the known data serves as the input and the projected outcome O_t as the output. The memory unit consists of three gates, denoted by green circles : the input gate, the forget gate, and the output gate. Furthermore, the cell's status is indicated by S_t, with the preprocessed data X_t and the previous state of the memory cell S_{t-1} serving as inputs to each gate.

The confluence points denoted by blue dots in Figure 3.3 represent multiplications, while the dashed lines symbolize the influence of the previous state. Observing the information flow within the memory unit's architecture, we can summarize the state update and output of the memory unit as follows (205) :

$$
\begin{aligned}
i_t &= \sigma\left(\boldsymbol{W}^{(i)} X_t + \boldsymbol{U}^{(i)} \boldsymbol{S}_{t-1}\right) \\
f_t &= \sigma\left(\boldsymbol{W}^{(f)} X_t + \boldsymbol{U}^{(f)} \boldsymbol{S}_{t-1}\right) \\
o_t &= \sigma\left(\boldsymbol{W}^{(o)} X_t + \boldsymbol{U}^{(o)} \boldsymbol{S}_{t-1}\right) \\
\widetilde{S}_t &= \tanh\left(\boldsymbol{W}^{(c)} X_t + \boldsymbol{U}^{(c)} \boldsymbol{S}_{t-1}\right) \\
S_t &= f_t \circ \boldsymbol{S}_{t-1} + i_t \circ \widetilde{S}_t \\
O_t &= o_t \circ \tanh\left(S_t\right)
\end{aligned}
$$

Here, ' $\circ$ ' indicates the Hadamard product, with i_t, f_t, and o_t representing the outputs of different gates. The new state of the memory cell is denoted by $\widetilde{S}_t$, S_t signifies the final state of the memory cell, and O_t represents the ultimate output of the memory unit. The coefficient matrices, $\boldsymbol{W}^{(i)}, \boldsymbol{W}^{(f)}, \boldsymbol{W}^{(o)}, \boldsymbol{W}^{(c)}, \boldsymbol{U}^{(i)}, \boldsymbol{U}^{(f)}, \boldsymbol{U}^{(o)}$, and $\boldsymbol{U}^{(c)}$, identified in Figure 3.3, play a crucial role. Through the operations of the distinct gates, LSTM memory units can effectively capture intricate correlation features within time series, both in the short and long term, showcasing a notable advancement compared to traditional RNNs.

The gating mechanisms within each LSTM cell empower it to make informed decisions regarding which information is pertinent to retain or forget at each time step. This makes LSTM well-suited for tasks such as time series forecasting, natural language processing, speech recognition, and various other applications requiring sequence analysis. Additionally, LSTM networks can be stacked, incorporating multiple LSTM layers interconnected with one another. This stacking approach enables the capture of hierarchical features and dependencies within the data, enhancing the network's capacity and its performance on intricate tasks.

LSTM became renowned for its capacity to effectively retain long-term dependencies. Nonetheless, the intricate architecture of LSTM neural networks often translates to prolonged solution times. In 2014, GRU emerged as a quicker training alternative, customized for machine translation due to its simpler design and easy implementation (37).

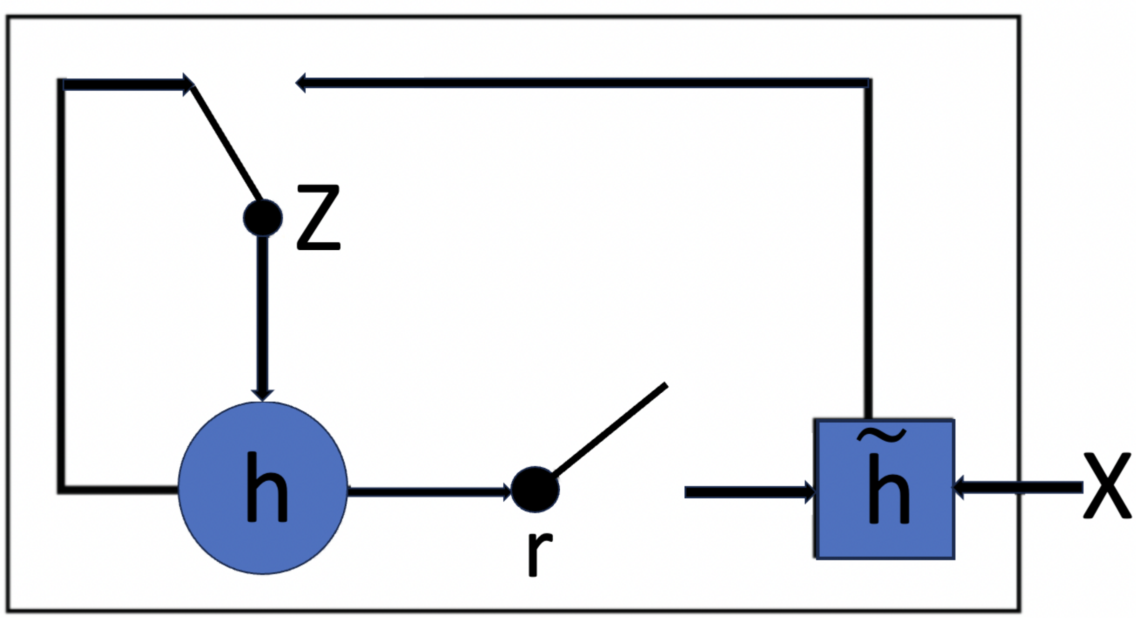


FIGURE 3.4 : GRU Cells Structure.

GRUs, or Gated Recurrent Units, present an alternative form of recurrent network that implements gating techniques to regulate information flow between cells within the neural network. GRUs share similarities with LSTMs but with a reduced parameter count. While featuring a reset gate and an update gate, GRUs lack the output gate. Consequently, the primary distinction between GRUs and LSTMs lies in the former's possession of two gates (reset and update gates) compared to the latter's three gates (input, output, and forget gates). This streamlined structure allows

GRUs to capture dependencies from extensive data sequences adaptively, preserving information from earlier segments of the sequence. As a result, GRUs are generally faster to compute, offering comparable performance (72). While GRUs have demonstrated superior performance on specific smaller and less frequent datasets, both variants of RNN have proven their efficacy in delivering the desired results. The standard configuration of GRU cells is displayed in Fig.3.4

In a standard GRU cell, there exist two gates : the reset gate (r) and the update gate (z). Much like the LSTM cell, the calculation of the hidden state output at time t involves the prior hidden state and the input time series value at time t, denoted in the equation.3.1 .

$$h_t = f(h_{t-1}, X_t) \tag{3.1}$$

The role of reset gates in GRUs resembles that of forget gates in LSTMs. Given the numerous similarities between GRU and LSTM neural networks, we won't delve extensively into the intricate formulas. Those interested in exploring this further can refer to (72) for more detailed information.

To address CNN limitations, Graph Neural Networks (GNNs) were introduced. They model network traffic using a graph representation, achieving a 16% lower MAE(Mean Absolute Error) compared to LSTM(172). GNNs factor in mobility when predicting events by removing low-weight edges. Traffic prediction models often require re-training for new or changed scenarios. An auto-encoder-based model learns compact BS representations from raw data, reducing computational costs and enhancing generalization(171). The model includes an encoder, spatial adder, and decoder, enabling it to infer representations that consider spatial relations with neighboring BSs. Numerical results demonstrate that this approach allows temporal models to achieve performance similar to spatio-temporal models, with a minor training time increase.

Transformer : The Transformer architecture represents a revolutionary neural network framework introduced in the paper titled "Attention Is All You Need" by Vaswani et al. In contrast to LSTM, which relies on sequential processing, Transformers pivot on the concept of self-attention mechanisms. This innovative approach allows Transformers to concurrently analyze all elements of an input sequence, endowing them with exceptional parallel processing capabilities suitable for sequences of varying lengths (short or long). Transformers have played a pivotal role in the success of numerous natural language processing tasks, including machine translation and language modeling.

The Transformer architecture is characterized by its encoder-decoder structure, making it especially suitable for sequence-to-sequence tasks.

1. **Encoder :**

   Input : Receives an input sequence, such as a sentence in a language.

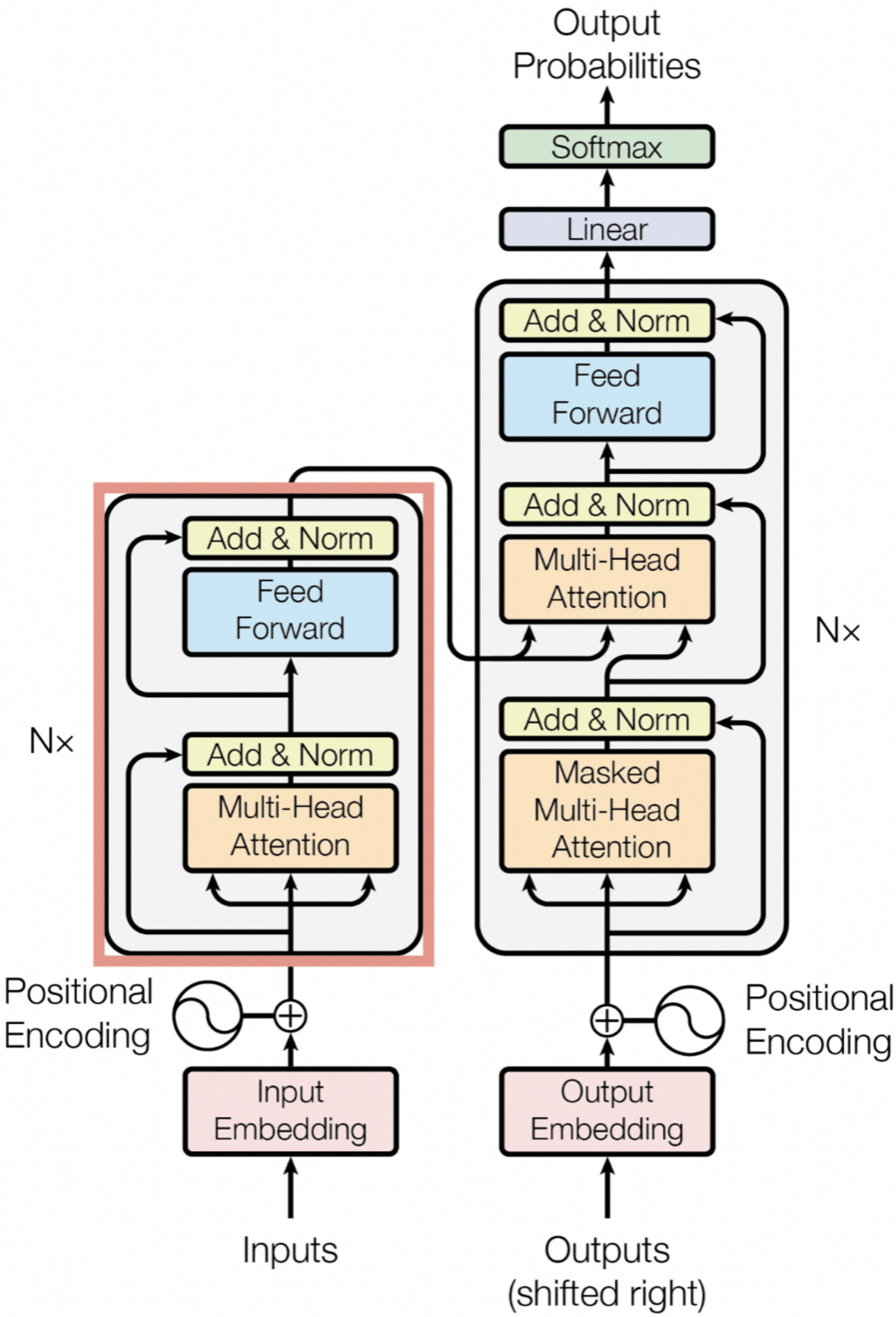


FIGURE 3.5 : The Transformer architecture's encoder-decoder structure, as presented in the paper "Attention Is All You Need".

Processing : The input goes through multiple layers, each comprising two key components :

— **Self-Attention Mechanism :** Aids in understanding the connections between words in the input sequence.
— **Feed-Forward Network :** Further processes the data.

Output : Produces an encoded representation of the input sequence.

2. **Decoder (for generation tasks) :**

Input : Utilizes the encoded representation of the input sequence.

Processing : Similar to the encoder but with some distinctions, including :

— **Masked Self-Attention Mechanism :** Ensures the decoder only considers preceding positions during decoding.
— **Encoder-Decoder Attention Mechanism :** Assists the decoder in concentrating on pertinent portions of the input sequence.

Output : Generates an output sequence.

The encoder handles input processing, and the decoder produces the corresponding output, demonstrating exceptional effectiveness in sequence-to-sequence applications as demonstrated in figure 3.5.

3.3 Conclusion

Within the context of future generation networks, machine learning has become an integral component of AI. Its integration is instrumental in effectively organizing and regulating network resources, providing efficient solutions for a range of technical complexities in the next-generation systems. This encompasses the management of device-to-device (D2D) communication, facilitation of large-scale massive MIMO, and the efficient administration of heterogeneous networks equipped with diverse technologies and architectures. In the pursuit of Energy Efficiency (EE) within modern networks, dynamic radio cell operation serves a pivotal role. This involves optimizing radio resource allocation through adaptable measures such as cell activation/deactivation and cell zooming to cater to fluctuating traffic demands. A comprehensive understanding of evolving user traffic patterns and mobility behaviors over time is critical for effective implementation. To further enhance predictive capabilities, the integration of advanced techniques such as deep Reinforcement Learning, including LSTM and GRU models, enables more accurate and efficient traffic predictions.

4 Power Model

4.1 Introduction

Studies have indicated that over 80% of the total power consumption in cellular networks is attributed to the radio access equipment, particularly Base Stations (BSs) (114). Certainly, a base station consumes a specific amount of energy to sustain its regular operation, encompassing energy usage for its circuits, cooling system, and other aspects. The objective of this research is to minimize the energy consumption of the Heterogeneous network. In this case, by modeling the electrical consumption of the entire network, we gain a better understanding of which components consume more energy than others, which unit is most affected by resources, and what the impact on the network is in terms of energy efficiency. Therefore, within this chapter, we assess the power model for each individual component and better control the activation/deactivation strategy.

4.2 BS Power Consumption

All BSs exhibit both dynamic and constant electrical energy usage patterns. The dynamic power is allocated to signal transmission, while the constant portion supports various operational functions such as signal processing, cooling, power supply, and backup battery charging (15). It's worth noting that the constant and dynamic power of a BS are interdependent. The dynamic power of a BS significantly impacts the power requirements of components like power amplifiers and cooling systems. For instance, research by Fehske (55) demonstrated that reducing a BS's transmit power from 20W to 10W results in a reduction in the BS's electrical power consumption from 766W to 532W. Figure 4.1 (40) illustrates the power consumption profiles of different sections within BSs.

To discern the right radio architectures that allow such a reduction in power consumption, studies have indicated that The average power usage of various components within today's wireless networks is significantly less compared to the power consumption of the base station component(48), which clearly shows that reducing the power consumption of the base station has to be an important element of our concern. The base station's overall efficiency, concerning the power it takes from its suppliers relative to its radio frequency (RF) power output, is controlled by the power consumption of its diverse components, including the core radio equipment. Fig.4.2 describes a block diagram of a complete base station with three sectors that can be generalized to all BS types, including macro, micro, pico and femto BSs. The antennas necessitate four transmit chains, resulting in a requirement of 12 power amplifiers (PA) per base station.To simplify the illustration, only one of these 12 transmit chains is depicted in Fig.4.2(75).

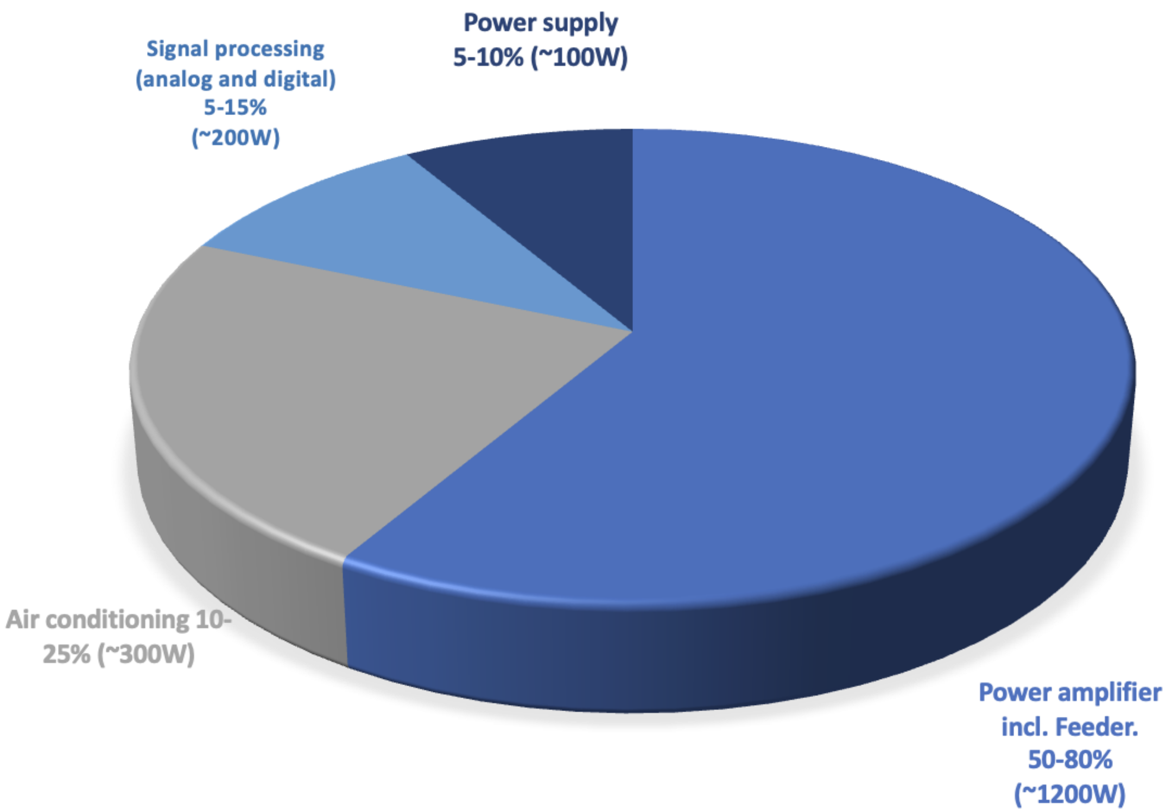


FIGURE 4.1 : Power Consumption Across Different Base Station Components.

The base station consists of the radio part (RF) whose power consumption corresponds to the amplification components (these devices amplify the transmit signals from the transceiver to a high enough power level for transmission, typically around 5–10 W), and radio transceiver (the equipment for generating transmit signals to and decoding signals from mobile terminals). Indeed, the component that consumes the most is the amplification (161) which depends on the output power, and itself depends on the needs of the UE-BS distance and the data rate of this user. It is relevant to note that the power amplifier efficiency for the macro-cell has been assumed to be 45%, which is in accordance with recently reported efficiencies (35% to 65%) for Doherty PA architectures, with advanced signal conditioning algorithms, performing at peak load (40), (179).

Considering the HetNet architecture, the modeling of both the macro cell (MBS) and small cell will be the same. In (7), the author describes the network's electrical consumption using a parameterized model that takes into account various factors such as bandwidth, the number of antennas, macro-BS sectors, and variable RRHs. However, regardless of the granularity of their model, this makes the on/off switching scheme much more complex. Additionally, they did not consider the reality of the network being heterogeneous, as we deploy both MBS and SBS base stations. Therefore, we provide a model that exclusively addresses this architecture.

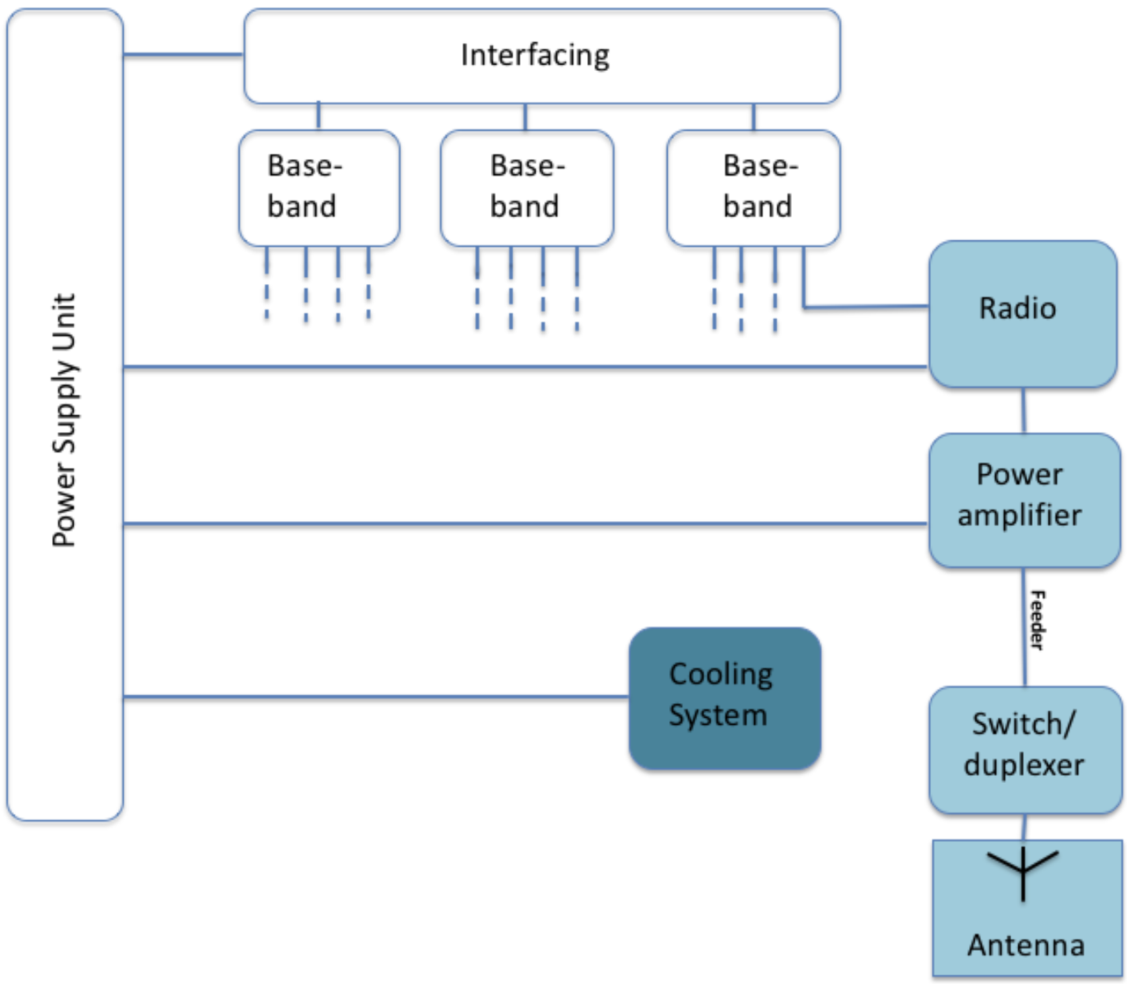


FIGURE 4.2 : Base station architecture configured for a three-sector system with four transmit antennas per sector, enabling MIMO functionality.

The BS is responsible for generating digital signals at the baseband (BB) level and then transmitting them to users through the RF transceiver. With the scaling factor technology, we can calculate the number of operations performed by the BS at the BB level per second per watt, or in operations per unit (GOPS)/watt, as shown in (4.1). In fact, (43) describes a generic power model that adapts to the base station hardware (BS type, number of antennas, maximum bandwidth, etc.) as well as its configuration (traffic load, energy-saving strategies, number of activated antennas, etc.). They explain how the scaling factor is used in the model. In essence, the power consumed by the components is derived from a reference power value defined by the supplier and combined with a set of scaling rules. This allows us to accurately measure the actual energy consumption when modifying a parameter such as bandwidth or the number of antennas. The document provides several tables of reference power consumption values and scaling factors. According to (43), there are six parameters that influence the energy consumption of any base station : bandwidth, spectral efficiency (which essentially includes coding rate and constellation order), number of antennas, frequency domain, and quantization.

$$P_{BB}=\sum_{i\in I}P_{(i,BB)}^{ref}\times\prod_{x\in X}\left(\frac{x_{ac}}{x_{ref}}\right)^{s_{i,x}} \tag{4.1}$$

Here, I represents the set of sub-components of the base station, and X is the set of parameters x discussed previously. $P_{(i,BB)}^{ref}$ is the power consumption of the components in baseband function load, while x_{ac} is the current value of power consumed by the parameters x. x_{ref} represents the reference power value, and $s_{i,x}$ is the scaling vector set to either 0 or 1. The latter indicates whether the power consumption of sub-component i is dependent on parameter x or not.

Furthermore, the base station requires an adequate continuous power supply to sustain its typical operation. Therefore, we should take into account the electrical consumption of the DC-DC conversion, which is formulated as follows :

$$P_{DC,DC}=\sum_{b\in M}l\left(\eta_{DC,DC}\right)P_{BB} \tag{4.2}$$

where $l\left(\eta_{DC},DC\right)$ is the loss function for DC-to-DC conversion efficiency, demonstrating that the converter's energy consumption is linearly proportional to the power consumption of the BS components, and M denotes the set of active BSs in the network. Furthermore, this DC voltage originates from an AC power source. Therefore, we must account for the energy consumption of AC-to-DC converters, which is expressed as :

$$P_{AC,DC}=\sum_{b\in M}l\left(\eta_{AC,DC}\right)\times\left(P_{DC,DC}+P_{BB}\right) \tag{4.3}$$

When all these components heat up, it becomes necessary to set up an active cooling unit for the BS hardware, the AC-DC converter, and the AC-AC converter. The latter is typically the unit that consumes the most energy, and it is proportional to the energy consumption of all the other components :

$$P_{\text{cool}}=\sum_{b\in M}l_{\text{cool}}\times\left(P_{AC,DC}+P_{DC,DC}+P_{BB}\right) \tag{4.4}$$

Therefore, the parameters of the base stations are already defined, and the hardware's energy consumption remains constant. Throughout the remainder of the thesis, we will consider it as such and assign it as P_{static} .

Furthermore, the base station also consists of the radio (RF) part, where the power consumption corresponds to the amplification components PA P_{PA} and RF radio P_{RF}. Indeed, the component that consumes the most is the amplification, which depends on the output power, itself is influenced by the requirements related to the distance between the UE and the BS as well as the data rate of the user, as indicated in reference (161).

In summary, we define the power consumption of a BS as follows :

$$P = P_{RF} \times N_{\text{antennas}} + P_{PA} + P_{\text{static}} \tag{4.5}$$

However, the amplification power (which is the power consumed by an amplifier), denoted as P_{PA} for base station b_i, depends on the base station's load and the amplification coefficient η_{PA}, according to the following equation :

$$P_{PA} = \frac{P_{tx} \times \rho_{b_i}}{\eta_{PA}} \tag{4.6}$$

where P_{tx} represents the transmission power per resource unit, and ρ_{b_i} is the number of resource units required to achieve the user's data rate. Let's assume that each UE k is served by b1. The user requires a data rate λ_k. And let's suppose that $d_{b_i,k}$ is the distance between the UE and BS, β is the path loss factor, and $h_{b_i,k}$ represents the fading effect. Therefore, the received data rate is calculated as follows :

$$r_{b_i,k} = B \times \log_2\left(1 + \frac{P_{tx} \times d_{b_i,k}^{\beta} \times h_{b_i,k}}{N_0 \times B + I}\right) \tag{4.7}$$

Given that N_0 is the thermal noise, I represents the interference, and B is the bandwidth. In order to guarantee good QoS by the base station, the following constraint must be ensured :

$$r_{b_i,k} \geq \lambda_k \tag{4.8}$$

To minimize the system's energy consumption, we require an optimal resource allocation at the BS level to avoid resource wastage. Document (129) introduced an algorithm for calculating the utilization rate and determining the optimal threshold at which activation/deactivation should be applied. In our case, we propose to apply a brute force search to find the optimal combination. According to the EARTH model, the electrical consumption of a BS in idle mode does not decrease because all the BS components remain powered on. Therefore, since we intend to power down the BS, it is only necessary to include the BS wakeup power. Some studies (130) and (43) consider wakeup power to be dynamic, proportional to the BS sleep period, which can range from a few microseconds to a few seconds. In fact, the longer the sleep time, the more components are deactivated, resulting in an increased need for wakeup power. Based on this, the total consumption of a BS b_i during a given period t is :

$$P_{b_i,t} = a_{b_i,t} \times P_{b_i} + |a_{b_i,t} - a_{b_i,t-1}| \times P_{sleep} \tag{4.9}$$

where

$$a_{b_i,t} = \begin{cases} 1, & \text{if base station } b_i \text{ was active during the period } t \\ 0, & \text{if base station } b_i \text{ was inactive during the period } t \end{cases} \tag{4.10}$$

Based on the above, Power consumption at BS in LTE can be categorized as static and dynamic power consumption :

— **Static Power Consumption :** Static power represents a fixed power consumption that is purely hardware-based and is required by the base station to support essential operations (and remains nearly constant). P_0 is the power consumption resulting from both site cooling and signal processing and occurred regardless of whether the BS is transmitting or not. Energy-efficient hardware designs and smart deployment strategies can mitigate this static power consumption.

— **Dynamic Power Consumption :** The dynamic power consumption (also known as communicational power) relies on the resource utilization of base stations and is directly influenced by their transmission operations. If the BS is transmitting, the transmit power P_{tx} is added to the total power consumption (98).

Therefore, we utilize the EARTH power model (18) in our simulator to estimate the power consumption of LTE system base stations. The relationship between radio frequency (RF) output power and base station power consumption exhibits almost a linear correlation, enabling a linear approximation. This power model is described by (4.11) (19) where N_{tr} stands for the total number of transceivers, P_{tx} is the relative RF output power, and δ is the sleep factor (Represents the cell DTX (Discontinuous Transmission) capacity in LTE networks, with $0 < \delta < 1$). The power consumed in the inactive state (sleep state) is constant and equal to δP_0, and P_0 Refers to power consumption when there is no load on the RF output power, where $P_0 \geq \delta P_0$. ΔP Denotes the slope parameter in the linear model for load-dependent power consumption (corresponds to how many RBs are transmitted). The ΔP slope is depicted in Figure 4.3 (165).

$$P = N_{tr} \times \begin{cases} P_{tx} + P_0, & \text{if transmission} \\ P_0, & \text{if no data transmission} \\ \delta P_0, & \text{if sleep mode} \end{cases} \tag{4.11}$$

As expected, The power consumption differs depending on the type of subframe in an active subframe. In pinned subframes, there is some power consumption, even if no Resource Blocks

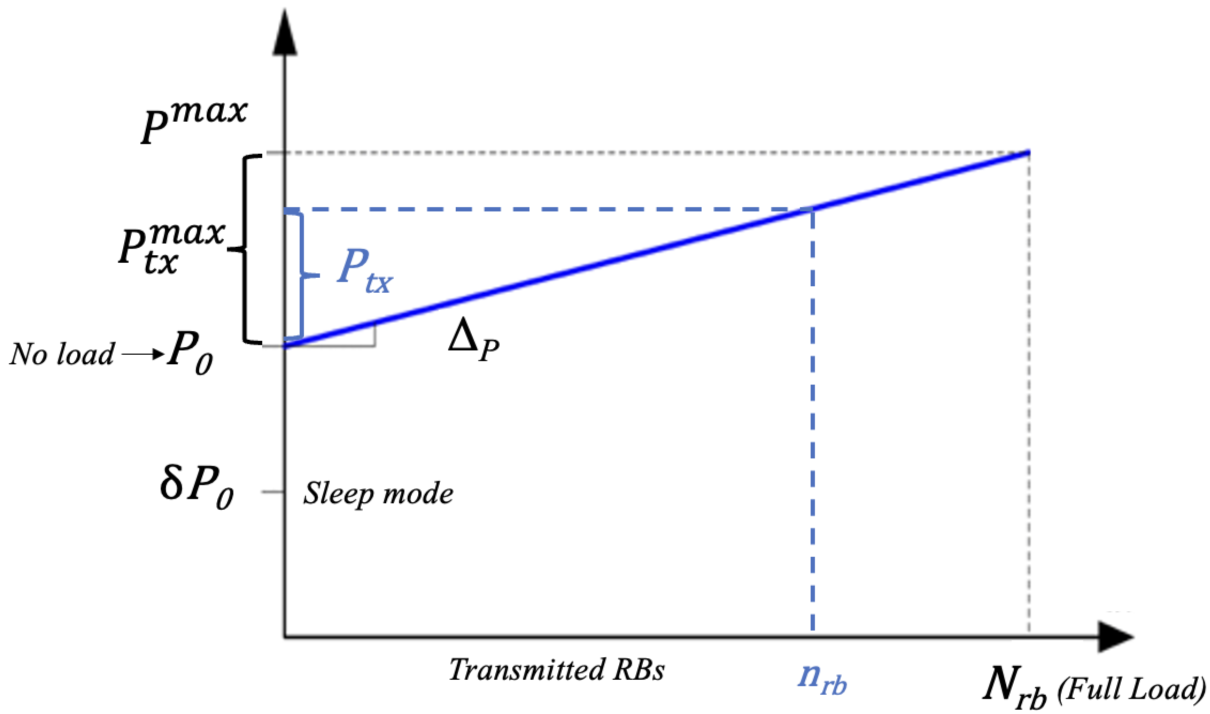


FIGURE 4.3 : PowerModel

(RBs) are allocated, for transmitting control channels like broadcast, paging, and synchronization. However, these channels do not need to be transmitted in free subframes. As a result, the baseline power P_0 will be greater for pinned subframes compared to free subframes. It's also evident that the maximum power consumed in an active frame (when N RBs are allocated) is independent of the presence of control channels and remains constant at P_{max}. Nonetheless, in the idle mode of base stations (when they have no active users within the cell and are not transmitting data), only mandatory reference signals are sent. During this phase, there is a total absence of user plane traffic, indicating that user data transmission is not taking place. In the case of LTE, all subframes transmit the CRSs (Cell-Specific Reference Signals).

We can conclude that the power consumption of the BS_ON (in transmission mode) is calculated as follows :

$$P = P_0 + P_{tx} \tag{4.12}$$

where the transmit power P_{tx} is calculated by the following equation :

$$P_{tx} = P_{rb} \times n_{rb} \tag{4.13}$$

n_{rb} denotes the number of RBs used. P_{rb} represents the power consumed by each RB (Resource Block) and calculated as follows :

$$P_{rb} = P_{tx}/N_{rb} \tag{4.14}$$

N_{rb} represents the total number of RBs.

4.3 Power Model in Sleep Modes

The power consumed during the sleep period corresponds to the power consumed by active components during that time. In fact, for the BS to be activated, it should not be completely turned off; it can still consume a certain amount of energy, such as detection power. In (130), they model the activation/deactivation and energy consumption during wake-up for all types of base stations based on the depth of sleep 1 (actual deactivation period) in which they define sleep modes in four states corresponding to the depth of deactivation : $71\,\eta\mathrm{s}, 1\ \mathrm{ms}, 10\ \mathrm{ms}$ et $1\ \mathrm{s}$. Their goal was to find the optimal deactivation duration corresponding to the traffic profile; this implies that deactivation and state changes should not exceed the period during which the base station is in sleep mode, while achieving energy reduction. This could be translated into constraints (4.17) and (4.18).

Let's assume that we have prior knowledge of the average daily traffic. Therefore, the traffic arrival process is a Poisson process with an arrival rate of λ and a service time of h_B. Additionally, the active time of the BS corresponds to the distribution of arrival times. Let's also assume that τ is the depth of deactivation in seconds, eq.(4.19). According to the document, the average energy consumption of the base station corresponds to :

$$E[P] = \frac{E\left[L_{\text{active}}\right] P_{\text{active}} + E\left[L_{\text{deac}}\right] P_{\text{deac}} + E\left[L_{\text{sleep}}\right] P_{\text{sleep}} + E\left[L_{\text{awake}}\right] P_{\text{awake}}}{L_{\text{active}} + L_{\text{deac}} + L_{\text{sleep}} + L_{\text{awake}}} \tag{4.15}$$

where $L_{\text{actif}}, L_{\text{désac}}, L_{\text{désactivation}}, L_{\text{awake}}$ correspond respectively to the active service duration, deactivation, deactivation time, and effective wake-up time. And according to the work (130), the power consumed during sleep is modeled as follows :

$$P_{\text{sleep}}(\tau) = P_m \exp\left(\left(-\omega_1 \log_{10}\left(\tau * 10^6\right)\right)^{\frac{1}{\omega_2}}\right) + \text{ constant} \tag{4.16}$$

P_{sleep} converges to the constant, which represents the power consumed at a certain sleep level. ω_1, and ω_2 are parameters that have values corresponding to different BS types.

In practical terms, the deactivation time should ideally be longer than the transition latency (4.17). Moreover, for optimal performance, the power consumed during the transition latency and sleep period should not surpass the power consumption in sleep mode (4.20).

$$\theta = \frac{L_{\text{awake}}}{L_{\text{awake}} + L_{\text{deac}}} > 1, \tag{4.17}$$

$$\eta = \frac{L_{\text{awake}}}{L_{\text{deac}}} > 1, \tag{4.18}$$

$$\tau = L_{\text{awake}} + L_{\text{deac}} + L_{\text{sleep}}, \tag{4.19}$$

$$P_{\text{deac}} + P_{\text{sleep}} + P_{\text{wakeup}} < P_{\text{idle}} \tag{4.20}$$

.

The numerical results indicate that if increased, they reduce the average energy consumption since they spend more time in sleep mode, resulting in only a marginal impact. However, if we increase the wake-up time relative to deactivation or sleep, the average power consumption increases. This suggests that we should exercise caution when selecting components to activate/deactivate when the sleep time falls within the configuration time range (τ and η).

GreenTouch has classified Sleep Modes (SMs) into four distinct levels, based on the grouping of sub-components with similar transition latency during activation and deactivation. The provided model allows for the quantification of power consumption in Base Stations (BSs) for each of these four SMs :

SM 1 : This level considers the shortest time unit, equivalent to one OFDM symbol (approximately 71μs), encompassing both deactivation and reactivation periods. In SM 1, only the power amplifier and a few processing components are deactivated.

SM 2 : SM 2 corresponds to a sub-frame or Transmission Time Interval (TTI) duration, which is around 1 ms. In this mode, a greater number of components enter the sleep state.

SM 3 : SM 3 aligns with the frame unit, spanning 10 ms. During this mode, most components are deactivated, resulting in reduced power consumption.

SM 4 : The deepest sleep level, SM 4, is characterized by units corresponding to the entire radio frame, which spans 1 second. This represents the standby mode where the BS is temporarily out of operation but retains the ability to wake up when needed.

Greater energy conservation can be attained by transitioning Base Stations (BSs) to deeper Sleep Modes (SMs) because a larger number of components are deactivated. Nevertheless, this shift comes at the cost of extended transition latency, which could potentially affect the Quality of Service (QoS) within the network. Table 4.1 provides an overview of the characteristics of the different SM levels.

TABLEAU 4.1 : BS sleep mode levels (43)

Sleep level	Deactivation duration	Minimum sleep duration	Activation duration
SM1	35.5 μs	71 μs	35.5 μs
SM2	0.5 ms	1 ms	0.5 ms
SM3	5 ms	10 ms	5 ms
SM4	0.5 s	1 s	0.5 s

However, in reality, and based on a cellular network traffic profile, the collected traffic data is averaged on an hourly basis. Thus, in our case, an idle period or a period of low traffic extends over hours. Consequently, in accordance with (4.15), the energy consumption variations during sleep mode can be considered negligible.

4.4 Energy Efficiency Metrics

Energy efficiency,in its conventional definition, is determined by the ratio of the total transmitted information to the total power consumption, denoted as Bit-per-Joule. This straightforward metric has been widely utilized in academic research, as indicated by references [(147),(83),(54),(74),(13),(25),(146),(202),(106),(101)], owing to its simplicity. However, with the rise of 5G technologies, several alternative metrics have come to the forefront. We distinguish between two types of communication scenarios : single-link and network-link. In the case of single-link communication, the energy efficiency metric is calculated as the ratio between the energy cost incurred and the benefits obtained after incurring this cost, as illustrated in (4.21). The academic community has explored various benefit functions, including system capacity (achievable rate), throughput, and outage capacity, among others.

$$\mathrm{EE} = \frac{\text{benefit}}{\text{energy consumption}} [bits/Joule] \tag{4.21}$$

Metrics for measuring energy consumption and performance trade-offs in energy-efficient design are classified into three primary categories : component-level metrics, equipment or node-level metrics, and system or network-level metrics (114). Component-level metrics assess the performance of individual wireless communication elements such as power supplies, power amplifiers, antennas, and others (12). On the other hand, equipment-level metrics measure the performance of network access nodes like User Equipment (UEs) and Base Stations (BSs). Network-level metrics evaluate the performance of network access nodes with regard to coverage area and expected Quality of Service (QoS) (131). Table 4.2 outlines commonly utilized metrics in this context.

TABLEAU 4.2 : Sets of energy efficiency measures

Measurement Level	EE Metrics	References
Component, Node, Network	bit/Joule	[(147),(83),(54),(74),(13),(25),(146),(202),(106),(101)]
Node	Energy Consumption Index (ECI)	[(121),(153),(91),(92)]
Network	Global Energy Efficiency (GEE)	[(123),(81),(34),(79),(124),(167),(193)]
Component, Node, Network	bit/Joule/Hertz	[(177),(125),(200),(30),(196),(195)]
Node	Weighted Product Energy Efficiency (WPEE)	[(167),(29)]
Network,Node	Area Power Consumption(APC) W/m2	[(121),(145),(143),(137)]
Node	Weighted Sum Energy Efficiency (WSEE)	[(167),(80)]
Component, Node, Network	Energy Consumption Gain (ECG)	[(22),(157)]
Node	Weighted Minimum Energy Efficiency (WMEE)	[(194),(47)]
Node	Energy Reduction Gain (ERG)	[(91)]
Network	Area Green Efficiency (AGE)	[(22),(157)]
Component, Node, Network	Absolute Energy Efficiency (AEE)	[(135)]

4.5 Conclusion

In this chapter, we have defined and presented the prerequisites necessary for the continuation of our study. In the following chapter, we introduce the brute force algorithm that inspired our strategy for activation/deactivation of base stations (BSs) with new proposed metric. In addition, we present a detailed analysis of our deep learning-driven switching control model, specifically designed to enhance energy efficiency within 5G networks.

5 Contributions

An in-depth examination of energy consumption in mobile cellular networks reveals that approximately 80% of the total energy is consumed by base stations (BS)(82). To address the foreseen surge in mobile traffic, 5G networks are taking steps towards extensive densification, placing a strong emphasis on the deployment of micro and pico base stations (BSs). However, the proliferation of BSs and the potential addition of antennas at each BS (ultra-massive MIMO) will significantly elevate network energy consumption, making the on/off switching of BSs a pivotal aspect of 5G.

However, the decision of when and which BS to deactivate is a crucial one, as it must ensure a certain level of quality of service (QoS) for end-users. In our approach, we propose leveraging an exhaustive search strategy to optimize the activation and deactivation of BSs. This strategy seeks to reduce energy consumption while considering the trade-off between power cost and QoS, adapting to varying user traffic patterns, thus enhancing overall system performance.

This section begins with an overview of the cellular HetNet utilized in our system model. We then delve into the proposed strategy, including inter-state transition costs and how to exploit it for formulating the optimal BS on/off switching strategy. Additionally, we provide an extensive analysis of our deep learning-based switching control model, meticulously designed to boost energy efficiency in 5G networks.

5.1 System Model

5.1.1 HetNet Architecture

5G network aims to establish full connectivity. And to meet this exponential growth demand in terms of capacity and throughput, the network has to upgrade its infrastructure. As a result, a dense deployment of several base stations with different coverage radius has proven to be effective.

To provide a brief overview of the system architecture. Figure 5.1 illustrates a cellular HetNet, served by M Macro BSs (MBSs) and N Small BSs (SBSs). The macro extends the network coverage to this area while SBSs (encompasses Femto, Pico, and Micro cells,) ensure additional localized capacity at the area edge and in hotspots. We also consider K non-uniformly distributed user-equipments (UE) changing their positions each TTI.

To calculate the received power, denoted as $Pr[dBm]$, the model that follows has been employed :

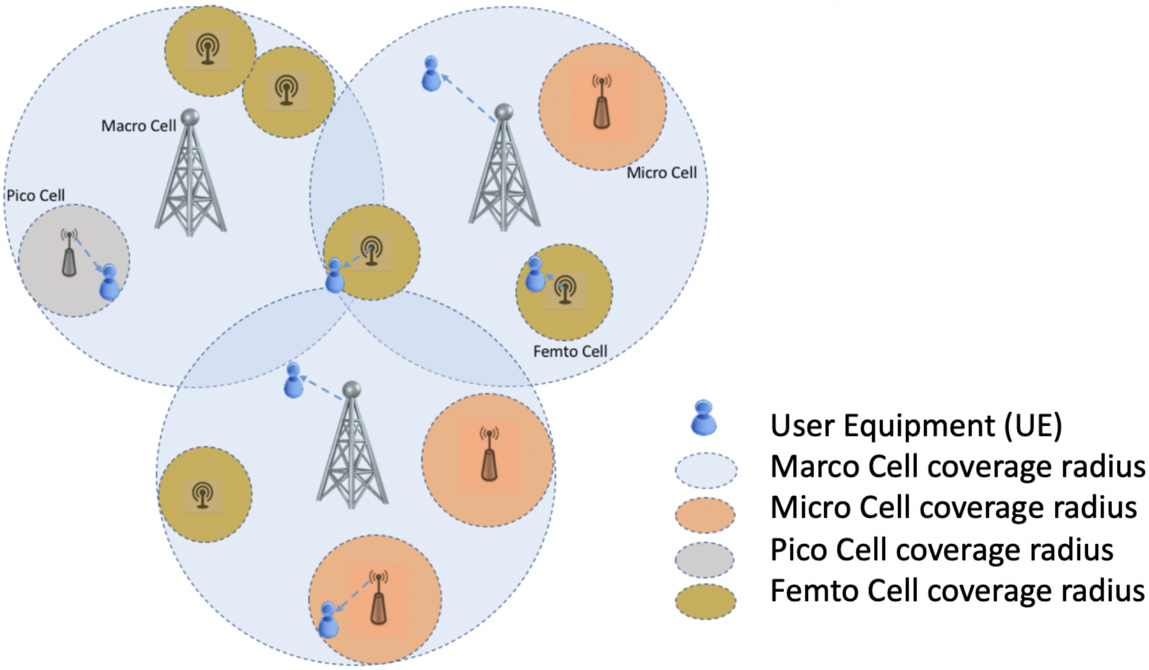


FIGURE 5.1 : HetNet Architecture.

$$P_r[dBm] = P_{tx}[dBm] + P_{Loss}[dB] \tag{5.1}$$

The received power P_r expressed in dBm is determined by considering the transmitted power P_{tx} also in dBm, along with the overall signal losses P_{Loss}. These losses are contingent on the given propagation area and are calculated as follows :

$$P_{Loss}[dB] = GA + PA \tag{5.2}$$

GA represent the overall gain of both antennas, and PA is the spatial transmission losses, which are computed in the following manner :

$$PA = \left(\frac{\lambda}{2 \times \pi \times d}\right)^{\eta} \tag{5.3}$$

Here d represents the distance to the BS, while η stands for the exponent loss,varying randomly within the range of $[2, 4]$. The signal to interference plus noise ratio (SINR) for UE k, is calculated as :

$$SINR_k = \frac{P_{rx,i,\bar{k}}[W]}{\sum_{i=1}^{M+N} P_{rx,i,k}[W] - P_{rx,i,\bar{k}}[W] + P_n[W]} \tag{5.4}$$

$P_{rx,i,k}$ and $P_{rx,i,\bar{k}}$ denote, respectively, the total power of BS i and its useful payload-carrying fraction (i.e., power portion of the transmit data destined solely to UE k), both received by UE k from BS i. The summation encompasses the total power received by UE k from all the BSs over the same frequency, and P_n signifies the noise power calculated in the subsequent manner :

$$P_n = -174 + 10 \log_{10}(BW) \tag{5.5}$$

where BW is the bandwidth of BS.

We define $\boldsymbol{b} = [b_1, b_2, \ldots, b_{M+N}]$ as the vector that characterizes the functioning states of the BSs, where

$$b_i = \begin{cases} 0 & \text{if the } i\text{-th BS is OFF} \\ 1 & \text{if the } i\text{-th BS is ON} \end{cases} \tag{5.6}$$

5.1.2 Proposed BS On/Off Switching Strategy

To figure out our multidimensional optimization problem. We need to address some challenges : The first is to develop a new metric that accounts for both the network energy cost and UEs' QoS. Our BS on/off switching strategy will rely on such a metric to make the appropriate choices. If the latter accounts only for the energy cost, it will result optimally in an all-BSs-Off configuration which would generate both a bit rate and a total utility null and, hence, a maximum of an optimization impossible to obtain. On the other hand, if the metric accounts only for QoS, it will result in an all-BSs-On configuration which is the most expensive in terms of energy. These extreme cases highlight the incontestable need for a metric that includes both the network's QoS and the energy cost.
The second challenge pertains to the fact that our BS switching strategy must exploit both instantaneous and predicted loads information. The instantaneous load is very useful to cope with the current network situation while predictions are keys to avoiding the bad impact of the current decision on the future network condition. To this end, we propose to start with exhaustive research which aims to optimize the numbers of activated and deactivated BS in a heterogeneous network. After building the approach to follow, let's now turn our attention to the cost of transitioning between states and how to exploit it to implement the optimal BS on/off switching strategy.

5.1.3 Inter-state Transitions Cost : Proposed Metric

As mentioned above, our metric U_T must account for both power cost (U_P) and QoS (U_Q) to ensure a trade-off between the user-perceived experience (i.e., network performance) and the incurred OPEX cost. Since we are operating an approach reflecting the network conditions at a different time (i.e. from present to near or even far future), this metric must also be dynamic (i.e. varying over time). Also, it must integrate any change in constraints and/or parameters as they evolve or appear over time. In this context, we have designed the following inter-state transition cost or metric :

$$U_T = (U_P)^w \times (U_Q)^{1-w} \tag{5.7}$$

U_T denotes the utility function to be maximized. w is the power cost weight. Note that w in (5.7) governs the QoS/cost trade-off to meet the operator policies. $w = 1$ or $w = 0$ emphasizes the power or QoS gain, respectively. In such a case, the decisions are made regardless of the QoS or the power cost. This means that the framework resulting from $w = 1$ or $w = 0$ is either all-BSs-off or all-BSs- on frameworks. The activation/deactivation decision takes into account both the traffic load of a BS and the quality of service of the UE, and generates the re-association of users, the recalculation of interference and finally the calculation of total consumption. Besides, to ensure energy efficiency and network throughput, we must simultaneously limit interference between cells.

The power gain U_P is measured by the following equation :

$$U_P = \frac{P_T(\mathbf{b}_0) - P_T(\mathbf{b})}{P_T(\mathbf{b}_0)} \tag{5.8}$$

$\mathbf{b}_0 = [1, 1, ..., 1]$ is the $M + N$ dimensional vector defined in (5.6) (i.e., b_i = 1 for $i = 1$, ..., $M + N$) characterizing the initial functioning states of the BSs when they are all ON i.e., b_i = 1 in (5.6) for $i = 1$, ..., $M + N$. P_T is the total power consumed by the group of active BSs given by :

$$\begin{aligned} P_T(\mathbf{b}) = N_{tr}^1 \sum_{l=1}^{M} \left(b_l \left(P_0^1 + P_{tx}^1 \right) + (1 - b_l)\, P_{sleep}^1 \right) + \\ N_{tr}^2 \sum_{l=M+1}^{M+N} \left(b_l \left(P_0^2 + P_{tx}^2 \right) + (1 - b_l)\, P_{sleep}^2 \right) \end{aligned} \tag{5.9}$$

where P_0^j and P_{tx}^j stand for the operation and transmission powers costs relative to the l-th BS in state i, respectively, N_{tr}^j is the number of tranceivers, P_{sleep}^j is the power associated to deactivate BS for $j \in \{1, 2\}$ (i.e., $j = 1$ for MBSs and $j = 2$ for SBSs), and b_l is a binary variable that takes the values 1 or 0 when the l-th BS is "on" or "off", respectively.

The QoS gain U_Q is calculated as:

$$U_Q = \frac{\Gamma(\mathbf{b})}{\Gamma(\mathbf{b}_0)} \tag{5.10}$$

$\Gamma(\mathbf{b})$ and $\Gamma(\mathbf{b}_0)$ are the average user throughput provided by active stations during a combination (the composition of on/off BSs) and the maximum throughput (the initial average throughput provided by the network when all BS are in ON state), respectively,

Highlighting the inherent flexibility and adaptability of our metric(5.7), we can further amplify its efficacy by introducing supplementary parameters, such as the total number of users served (who are benefiting from the network). This makes our metric even more impressive and versatile, as it can handle a wide range of different situations by seamlessly integrating additional factors. This adaptability means that we can easily adjust when to turn on or off the BS units based on what the network operator specifically wants. Whether the goal is to save power, increase data (in terms of throughput gain), or give the users the best experience by considering how many of them are being served –our strategy remains super flexible, able to meet the different goals of managing the network. So, our metric goes beyond just being a calculation; it grows into a tool that captures the impressive ability to adapt and handle the complexities of today's communication systems.

To improve the quality of experience (QoE) in our approach of BS switched on/off, we have chosen to introduce a new constraint in the calculation of the metric(5.7) which is represented by the number of served users :

$$U_T = (U_P)^{w_1} \times (U_Q)^{w_2} \times (U_C)^{w_3} \tag{5.11}$$

where U_C is the ratio of served UEs. w_1, w_2, w_3 are the power cost weight where $w_1 + w_2 + w_3 = 1$. Note that $w1, w2, w3$ can be modified according to the operator policies (i.e., favoring either the power, throughput gain (based on link quality), or user experience (based on number of users served)) for giving more weight to the Cost, QoS or user experience.

Accordingly, the optimal on/off states can also correspond to the case that equalities hold for the constraints proposed in the metric calculated using equation(5.11).

5.2 Brute Force (BF) Algorithm

Starting with a brute force (BF), also referred as exhaustive Search (ES) algorithms, in problem-solving offers a foundational strategy that aids in comprehending the intricacies of a problem. By meticulously exploring all possible configurations, it establishes a clear benchmark for evaluating the efficiency gains achievable through the subsequent optimized algorithms. This initial method serves as a practical groundwork, demonstrating the feasibility of the switching process and potentially uncovering unexpected complexities. Moreover, it facilitates a clear comprehension of the network's behavior and ensures a solid understanding of the problem's nuances before embarking on complex optimizations.

Initiating the process with a brute force approach in the context of base station activation and deactivation, while incorporating a utility function to balance energy consumption and throughput gains for Quality of Service (QoS) assurance, serves as a foundational step towards effective network management. By exhaustively analyzing various combinations of base station states, the brute force strategy establishes a performance baseline that aids in evaluating the effectiveness and offers the best possible solution for cell switching. Introducing the utility function, which suggests the delicate balance between energy efficiency and network performance, helps into the decision-making process. This initial stage not only validates the capability of managing base stations for optimal QoS but also provides a real understanding of the interplay between energy considerations and data throughput.

In this regard, the purpose is to find a partition that groups together the most well-used base stations and minimizes the sum of energy consumption while maintaining the quality of service. Our metric (5.7) accounts for both power cost (conventional or green) and QoS where the trade-off between the user-perceived experience (i.e., network performance) and the incurred OPEX cost is factored according to the user traffic.

Nonetheless, with a growing number of base stations, the computational complexity escalates exponentially, the brute force approach becomes computationally intensive and impractical. Here, machine learning algorithms step in, leveraging historical data and patterns to predict optimal base station configurations. These algorithms adapt to evolving network conditions, optimizing QoS while minimizing energy consumption. Through predictive analytics, it can forecast traffic patterns and energy requirements, ultimately enhancing QoS while accounting for energy constraints. Thus, while a brute force approach clarifies fundamental dynamics, machine learning algorithms offer scalability and adaptability to handle the dynamic and data-intensive nature.

In our case, before adopting our machine learning-based approach for prediction, we initially used the 'brute force' method as a reference, considering its exhaustive nature and its ability to provide perfect solutions. The 'brute force' method is thus perceived as a practical benchmark, representing the limit of prediction accuracy in a context where an exhaustive solution is known.

However, to establish a connection between our deep learning approach and the theoretical foundations of statistical estimation, we introduce the concept of the Cramér-Rao Lower Bound (CRLB). The CRLB represents a theoretical limit on the accuracy achievable by any unbiased estimator in the field of statistical estimation. While our 'brute force' method acts as a practical benchmark, the CRLB provides a theoretical perspective on the best theoretically possible accuracy in a statistical framework.

Thus, considering the CRLB as a conceptual link, our choice to use the 'brute force' method as a practical benchmark to assess the performance of our deep learning approach is justified. When our prediction model achieves perfect accuracy, it approaches the performance of the 'brute force' method, reinforcing the validity of our approach in relation to the theoretical limits defined by the CRLB.

The synergy between initial brute force analysis and subsequent machine learning empowers efficient, adaptable decision-making in complex, dynamic telecommunications environments.

5.3 Deep Learning Switching Control

Despite offering an optimal converging solution, the brute force approach detailed in previous section necessitates excessive computational resources and time. Hence, we believe that employing deep learning (DL) represents an excellent resolution. This is because an adept neural network, trained on data generated through the previous algorithm, can yield comparable outcomes. Nevertheless, this process is expected to be less resource-intensive and faster in terms of computation.

5.3.1 Core Framework

In the realm of deep learning, an algorithm processes a dataset comprising input information and the corresponding anticipated output values. Through this process, it aims to comprehend the underlying patterns, allowing it to apply this understanding to new and unseen inputs. To uncover these patterns, deep learning utilizes Artificial Neural Networks (ANNs), which are constructed with interconnected nodes organized into layers including input, output, and hidden layers. These layers contain neurons responsible for producing output based on input values (x), weights (w), and bias (b). It can be mathematically described as $z(i) = w * x(i) + b$. Notably, the model refines these latter components during training. Usually, the Gradient Descent (GD) technique is employed, which involves updating the weights and bias through hundreds or thousands of iterations(73). In each iteration, these elements are updated in proportion to a learning rate, a parameter that governs the speed at which the algorithm learns, ultimately moving towards the desired outcome.

To determine the adequacy of the weights, the neural network forwards its inputs through to its outputs and assesses the predictions by comparing them to the expected values. This evaluation can involve utilizing a loss function such as "Mean Squared Error" or even more convoluted functions like the one provided below :

$$L(\hat{y}, y) = -y \times \log(\hat{y}) - (1 - y) \times \log(1 - \hat{y}) \tag{5.12}$$

Here, $\hat{y}$ represents the prediction and y signifies the expected value. The objective is to reduce this function's value by modifying the neural network's weights.

Epoch and Batch size : epochs are defined as complete passes through the training dataset, while the batch size determines the number of samples processed simultaneously through the network before the model is updated. An increase in the number of epochs can aid in reducing losses.

5.3.2 Strategic Design of Learning Framework

Undertaking a pioneering initiative to enhance energy efficiency in wireless networks, our research endeavors to employ deep learning techniques to train a pair of neural networks.

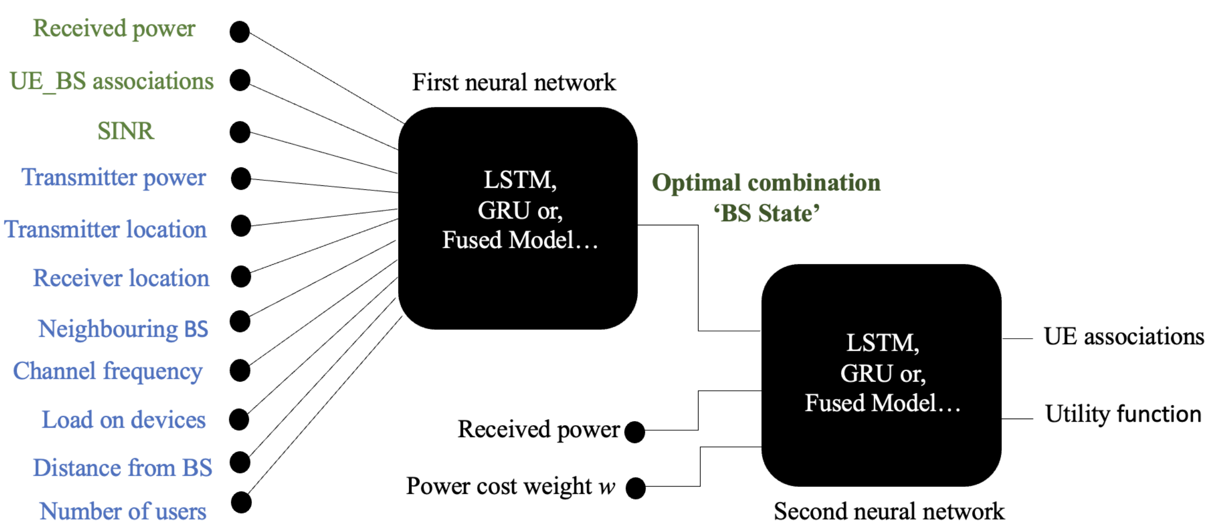


FIGURE 5.2 : System design.

The first neural possesses the capacity to incorporate a wide range of input parameters (including factors such as Number of users, Received Power, Distance from BS, Neighbouring BS, load en devices, Transmitter location...) to predict non-critical base stations and subsequently generate base station status outputs. This algorithm provides binary assessments for each individual base station, thereby categorizing their operational state as either 'ON' or 'OFF'.

The second neural uses the output of the first neural as input parameter to provide the user associations to the base stations offering a comprehensive resolution for enhancing energy efficiency within HetNet networks. This harmonious approach yields an all-encompassing solution that greatly contributes to the amplification of energy efficiency, ultimately resulting in a significant reduction in power consumption.

To construct our model, we employed the MATLAB programming software, utilizing our programmed 5G simulator. The schematic representation of our proposed model is depicted in Figure 5.2. The parameter values (mentioned in green color) chosen for constructing our first Neural Network as a first step. However, other parameters mentioned in blue will be tested in our future work for better performance

To facilitate the training of our model, we initially generated a dataset that incorporated diverse parameters, such as the number of users, received power, distance from BS, Neighbouring BS, transmitter power, etc. This dataset was created by running MATLAB code for parameter generation multiple times. Subsequently, the network, as described earlier, experienced training using this dataset. These generated parameters were then categorized into the optimal BS combination by using DL.

As we previously introduced the idea of incorporating a second neural network into our system, the initial neural network's results can provide a solid input parameter to the second neural network. The first neural helps identify which Base Stations (BS) remain active/inactive. Additionally, these results can further refine the accuracy of this second neural network. With this approach, our objective is to rely solely on deep learning to comprehensively configure the network. In essence, the output of this neural network will determine the associations between users and Base Stations (BS).

Regarding the configuration of our deep learning approach, our initial testing involved setting up the neural networks across three different models (LSTM, GRU, and the configured Fused model). The objective was to identify the optimal model that could enhance energy efficiency while preserving service quality.

The selection of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) as recurrent neural network (RNN) architectures for predicting off-base stations and enhancing energy efficiency in wireless networks is deliberate and distinguishes them from other machine learning models. Specifically designed to mitigate the vanishing gradient problem, LSTMs and GRUs excel in handling sequential tasks like time series prediction. LSTMs leverage a complex memory cell structure for capturing prolonged dependencies, while GRUs employ a simpler architecture with gating mechanisms, potentially leading to quicker training and improved performance on tasks with shorter dependencies. The preference for LSTM and GRU over other machine learning models stems from their innate ability to effectively model temporal dependencies, a critical factor in addressing the dynamic nature of wireless network data. This deliberate choice underscores the strategic decision to employ RNN architectures tailored for sequential data, setting them apart as more apt solutions for the intricacies involved in predicting off-base stations and optimizing energy efficiency in wireless networks.

The LSTM and GRU models were described in Chapter 3/Section 2 (for more comprehensive details, please refer to (59)), while the fused model will be elaborated on in the following section.

5.3.3 Fused Model Description : A Deep Multimodal Learning

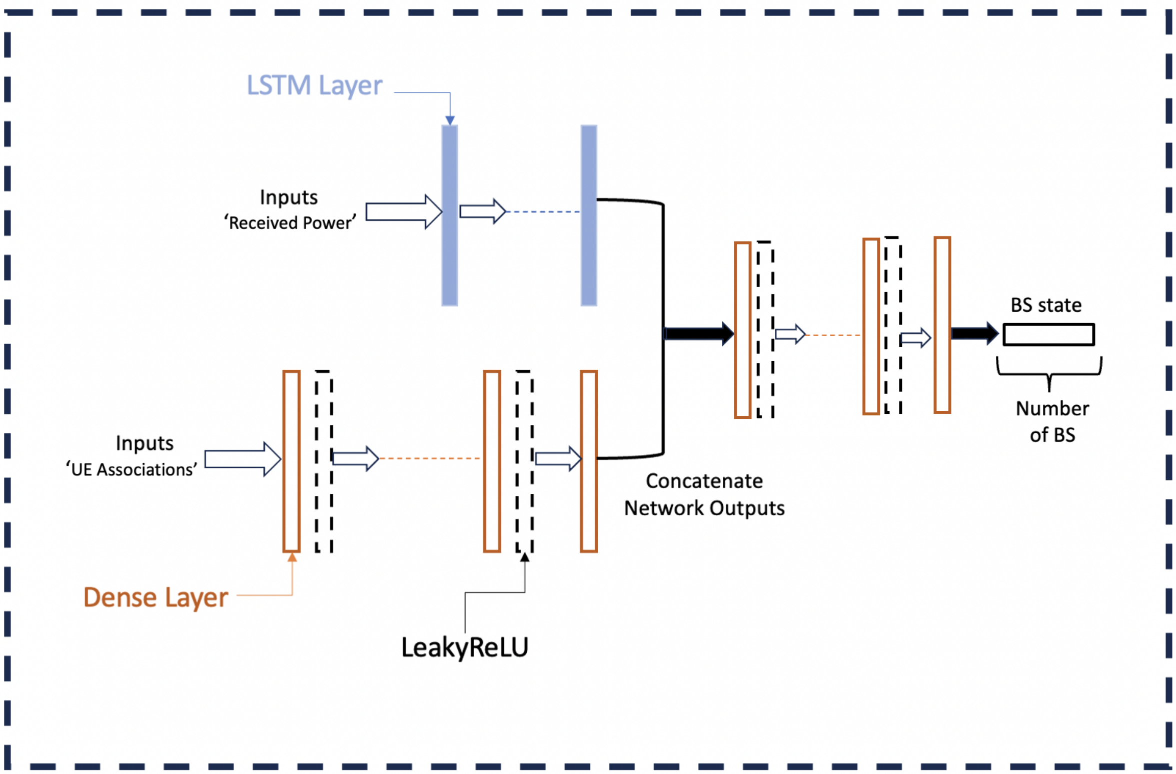


FIGURE 5.3 : Fused Model Structure.

To improve our system's performance, a deep learning multimodal model (DML) was constructed with inspiration drawn from the methodology presented in the article "Deep Multimodal Learning : Merging Sensory Data for Massive MIMO Channel Prediction" by Yang et al.(188). While we selectively adopted a segment of their model, we took a distinct approach in formulating a novel architecture. Illustrated in Figure 5.3, our model integrates a series of LSTM models with Dense layers, incorporating LeakyReLU activation functions, to effectively forecast the state of base stations. Notably, our approach leverages the received power as the input for the LSTM sequences, while the user association serves as the input for the Dense layer with LeakyReLU. This fusion of diverse data streams enables our model to capture intricate patterns and deliver accurate predictions in complex, real-world scenarios.

Comprising two networks, the fused model incorporates multiple LSTM (Long Short-Term Memory) layers in the first network, designed for analyzing time series data and capturing temporal patterns. The second network includes several dense layers, each employing the LeakyReLU function, with the exception of the output layer, which serves as the activation function. As shown in Fig.5.3, the concatenated network combines the outputs of the two models, which in turn consist of dense layers with the LeakyReLU function, effectively integrating these modalities at the decision levels.

The dense layer, also referred to as a fully connected layer, represents a fundamental type of neural network layer. It involves each neuron within the layer being connected to every neuron in both the previous and subsequent layers. This layer is characterized by its use of weights and biases to learn intricate patterns within the data during the training process. It is commonly used in various deep learning models for tasks such as classification, regression, and feature learning.

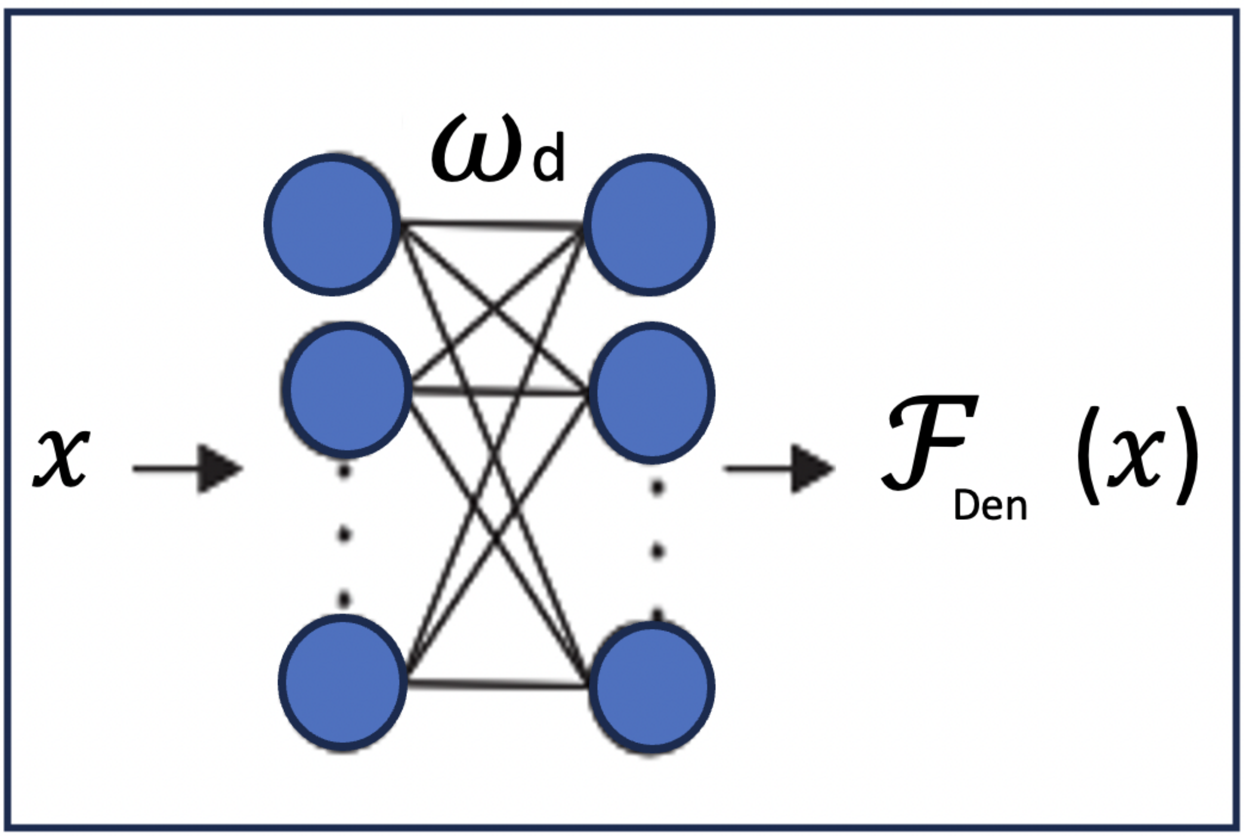


FIGURE 5.4 : Dense Layer Structure.

The mathematical representation of the dense layer is given by $\mathcal{F}_{\text{Den}}(\boldsymbol{x}) = \boldsymbol{w}_d\boldsymbol{x} + \boldsymbol{b}_d$, where $\boldsymbol{w}_d$ represents the weight and $\boldsymbol{b}_d$ denotes the bias associated with the dense layer.

In general, LSTM layers are specifically designed for handling time series data and capturing temporal dependencies, while dense layers with LeakyReLU activation functions are more general-purpose layers used for feature extraction and non-linear mapping in neural networks.

The LeakyReLU activation function is a variation of the rectified linear unit (ReLU) commonly integrated into neural networks. It operates by applying a non-linear transformation to the outputs of network layers. Mathematically, the LeakyReLU function is defined as $FLR(x) = max[x, 0.2x]$, where x represents the function's input. By introducing a slight slope (0.2) for negative values of x, it allows some information to flow even when the input is negative. This prevents the occurrence of the "dead ReLU" phenomenon, where neurons can become inactive, impeding learning. The implementation of the LeakyReLU activation function, instead of the standard ReLU, aims to enhance the neural network's performance and learning capabilities.

In summary, the fused model (FM) using LSTM and Dense layers, with dense layers followed by the LeakyReLU function, will be utilized to predict the combination of on-off base stations using power received and user associations as inputs by training the model on a dataset that includes the power received and user associations as input features and the on/off base station combinations as the target labels. The LSTM layers in the model can capture the temporal dependencies in the input data, while the Dense layers with LeakyReLU activation functions can introduce non-linearity and learn complex patterns in the data. By training the model on a sufficiently large and diverse dataset, it can learn to predict the on-off base station combinations based on the given inputs.

5.4 Conclusion

The results from the simulations demonstrate that architectures based on DML (Deep Multimodal Learning) offer notable advantages over those based on a single modality in most scenarios. These findings underscore the effectiveness of the proposed framework in improving energy efficiency by efficiently managing the ON/OFF switching of base stations in response to traffic fluctuations, thereby providing valuable support to current wireless communication systems.

6 Performance Analysis and Simulation Results

To evaluate the performance of our proposed algorithms, a simulation platform is developed, which defines the structure of our work, we will explain the established channel models, the planning of the association of users in our simulator as well as the parameters used for our simulations. Finally, we present the results and analyze the performance of our approach.

6.1 Simulator Description

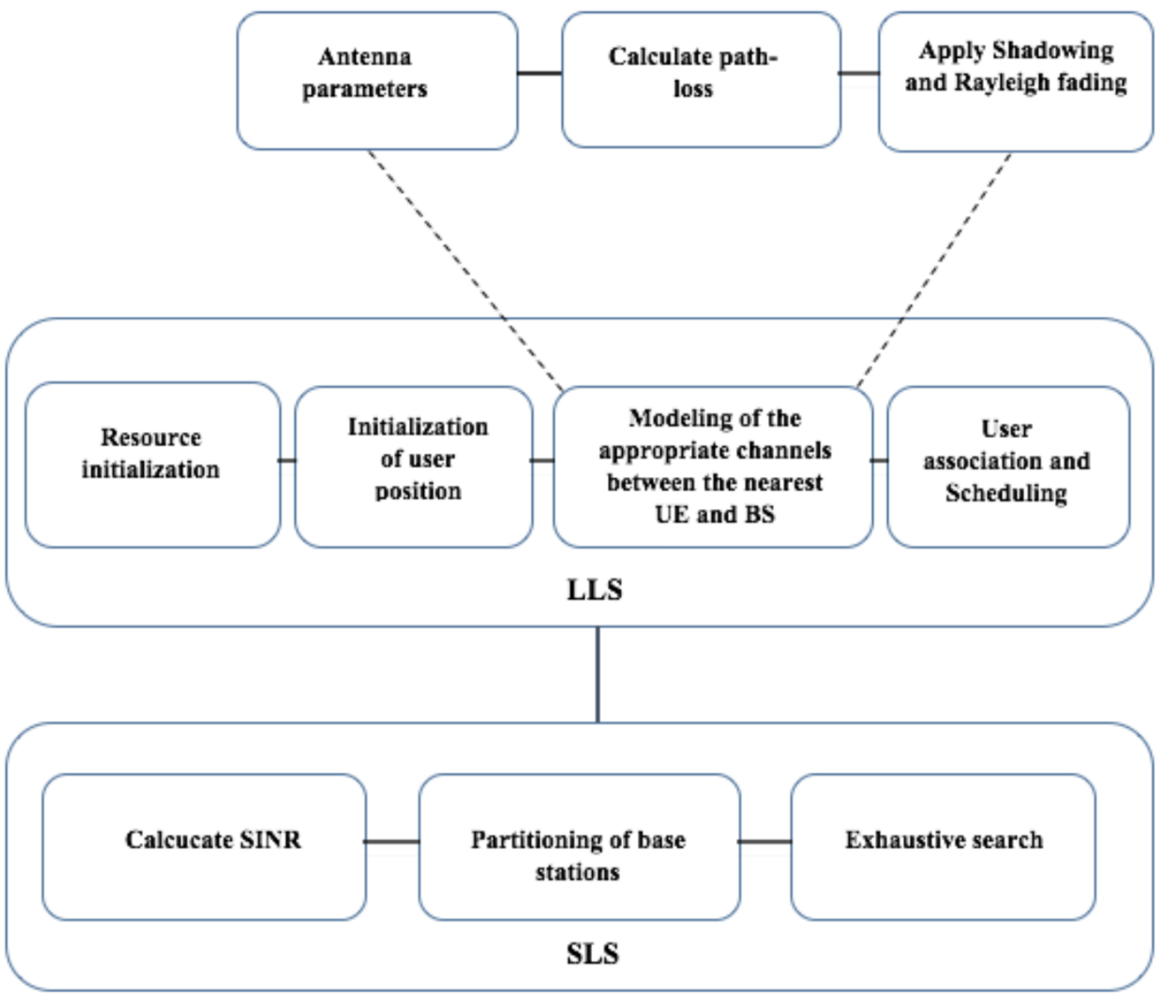


FIGURE 6.1 : Simulator Structure.

For our simulator, we opted for a simulator based on LTE as our concern is mainly based on the downlink signal and the fact that 5G uses the same signal as LTE. Our signal frame would be an

OFDM with 7 symbols and 1 ms TS (Time slot). In addition, as our network being heterogeneous, we are deploying micro and macro base stations to improve the quality of service for users. Thus, our simulator should take into account both environments.

To implement and deploy our network, we will simulate our local HetNet on two levels : Link level simulation (LLS) which concerns the radio channel between the transmitter antenna and the receiver antenna where we generate a channel model that covers the propagation of urban macro and micro-cells as well as system-level simulations (SLS) that represent the functionality between BS and UE. In this layer, we describe resource allocation, and user association which would provide us with enough data (such as SINR and throughput) to assess the performance of our network. Thus, our simulator will provide us with key performance indicators that we need for our energy efficiency approach.

Fig. 6.1 represents a graph describing the sequence of the construction process of our simulator. The measurement of the quality of the channel between the surrounding base stations and the users leads us to generate a channel at each TTI. To this end, we will calculate the quality of the channel for different TTI scenarios for more than one hour and average these results to have a statistical channel average.

Channel Model :

The channel model definition comprises two types of fading : large-scale fading (LSF) and small-scale fading (SSF). LSF occurs when an obstacle comes between the receiver and the transmitter. However, the signal power experiences significant fluctuations due to two components of LSF : path loss and shadowing. As for small-scale fading, the received signal undergoes rapid fluctuations over short durations and distances.

Our path loss model is based on the COST Hata propagation model, which is well-established for urban propagation scenarios in systems operating within frequency ranges up to 2GHz. Both path loss and shadowing effects depend on the user equipment's changing position over time. In contrast, small-scale fading is a time-dependent model.

In our simulator version, we will rely on the Rayleigh channel model for a two-antenna transmission mode, generating channel coefficients for each UE-BS connection. Nevertheless, the complexity of generating these coefficients for a single scenario grows exponentially with the total number of links between a user and all antennas. To mitigate this computational complexity, we will pre-generate channel coefficients solely for the N closest base station antennas to a user and create links as required.

To achieve a more detailed understanding of small-scale fading effects quickly and with reduced complexity, we can pre-generate channel coefficients for different delay and angle of arrival scenarios offline. This approach allows us to obtain a statistically realistic measurement of the channel's state during an entire hour by averaging these realizations.

TABLEAU 6.1 : Simulation Parameters

Parameters	**Macro BS**	**Pico BS**
Carrier frequency	2 GHz	2 GHz
Bandwidth	20 MHz	20 MHz
Maximum number of RBs N_{rb}	100	100
Transmission power	46 dBm	22 dBm
Coverage radius	1 Km	100 m
P_{base}	260 W	6.8 W
P_{sleep}	218.4 W	4.3 W
Number of antennas	2	2
Antenna height	30 m	6 m
Mobile user height	1.5 m	1.5 m
TTI	0.001 s	0.001 s

Tab. 6.1 summarizes all the parameters used in the simulation of network deployment. As described, we used an LTE simulator to generate macro-cells positioned according to a hexagonal architecture, then it randomly distributes Pico sites and users within each macro coverage region. During the simulation, each user can be attached to the cell providing the best RSRP (Reference signal Receive Power), but since the user moves at each TTI, we find ourselves in the case where the user requests a transfer periodically. To avoid this, however, we averaged the condition of the channel for one hour.

6.2 Simulation Results

6.2.1 Brute Force Algorithm Results

For the sole sake of simplicity, we started with a network of hexagonal architecture that deploys 3 MBSs and 2 SBSs, and 50 stationary UEs per MBS (i.e., 511 maximum number of combinations equals to all possible switching). The MBS is placed in the center of the hexagon, whereas the SBSs are randomly positioned as shown in the Fig. 6.2.

By analogy with our work, our approach aims to find the transformation (on/off) which would maximize the cost expressed by equation (5.7).

Fig. 6.3-a shows the cost function (5.7) for different values of the power weight w. This metric explains both the reduction in power and the gain in QoS. Indeed, a high value of w means that the power reduction prevails over quality of service. In this regard, we have launched the search

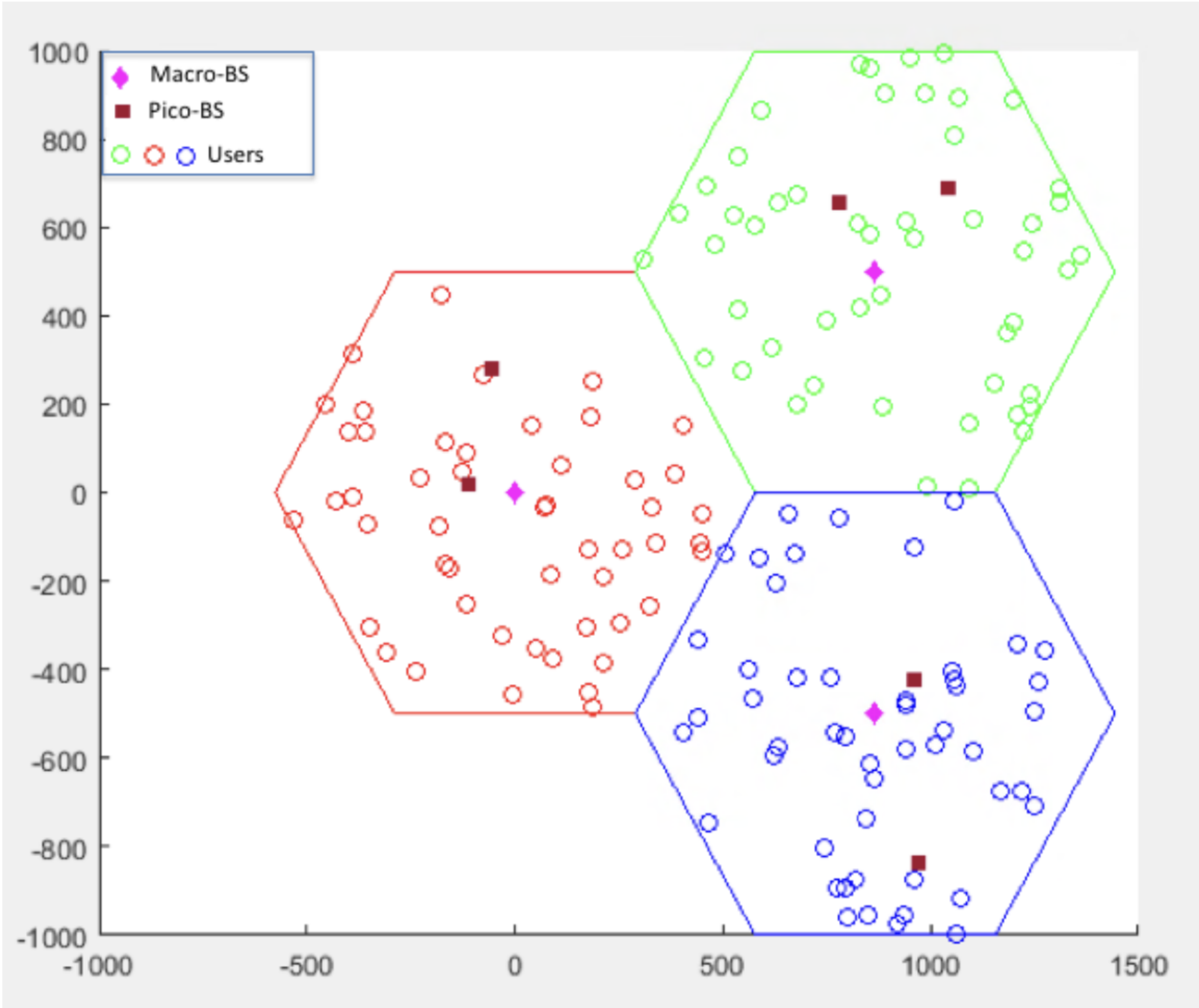
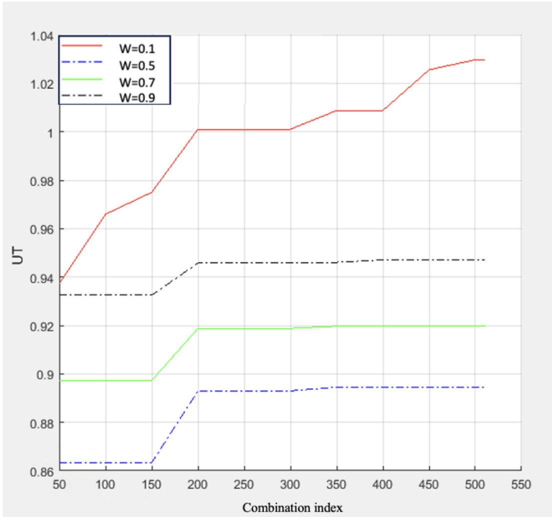


FIGURE 6.2 : Distribution of users in the cellular network

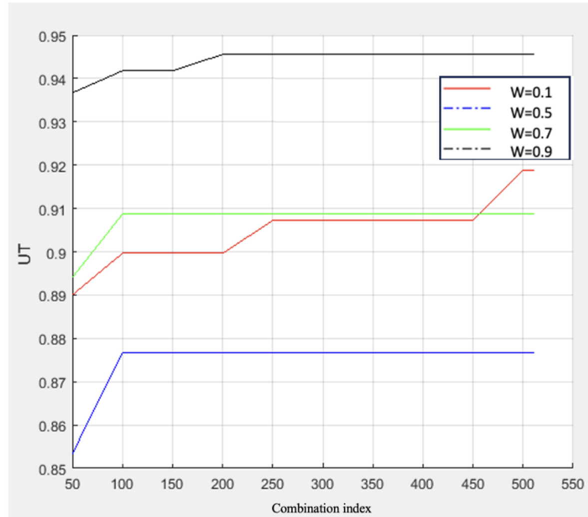
for each of the values of $w = 0.1$, $w = 0.5$, $w = 0.7$, and $w = 0.9$ in order to analyze the behavior of the search when power or QoS takes priority.

For the given values of $i\_max = [50 : 50 : 511]$, Figs. 6.3-a and 6.3-b represent the utility function U_T as a function of w. I_opt indicates the optimal number of iterations in the interval $I = [1,$i_max$]$ where we obtain the maximum value V-max of the utility function.

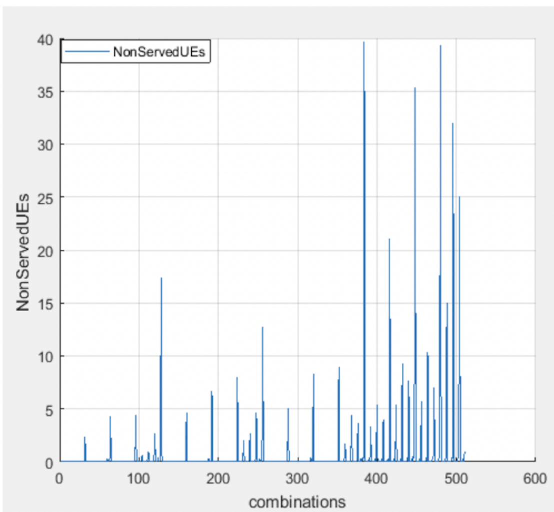
In Fig. 6.3-a, the maximum value of U_T (when $w = 0.1$) is recorded with only 2 MBS in the ON state, while for $w = 0.5/w = 0.7$ the maximal value is recorded with 3 SBSs at on state , and for $w = 0.9$ with the combination of 2 SBSs in on state. Equally in Fig. 6.3-b, the maximum value of U_T(when $w = 0.1$)is recorded with only 2 MBSs and 1 SBS in the on state, while for $w = 0.5/w = 0.7$ the maximal value is recorded with 4 SBSs at ON-state, and for $w = 0.9$ with the combination of 3 SBSs in ON-state. Therefore, it can be concluded that the number of BS in



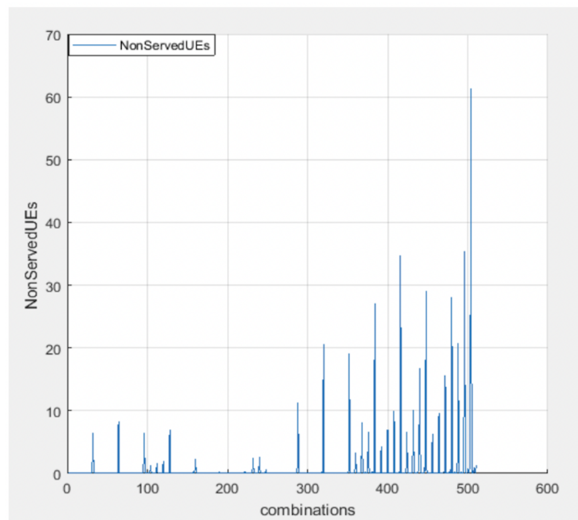
(a) Utility function (5.7),for several values of w.



(b) Utility function (5.7), for several values of w.



(c) Non served UEs.

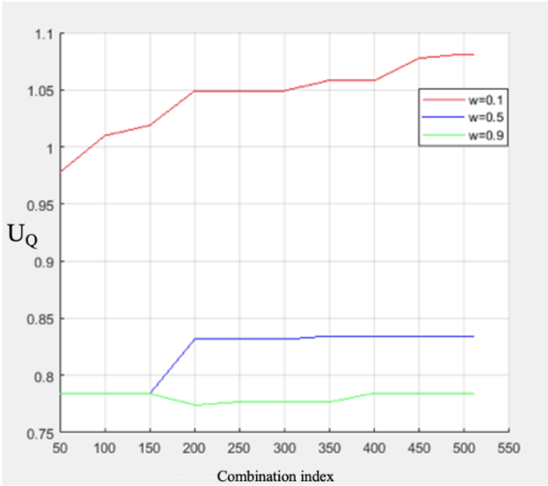


(d) Non served UEs.

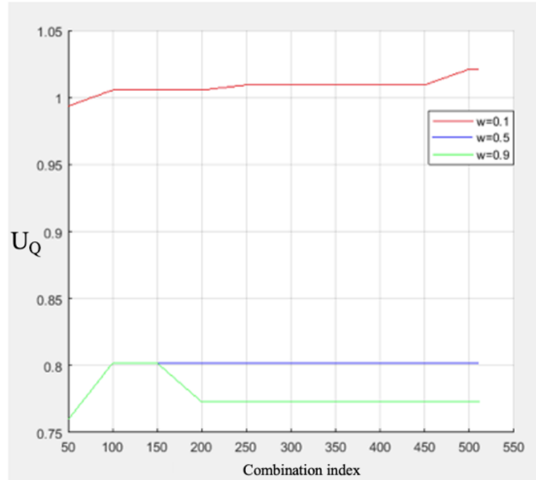
FIGURE 6.3 : (a), (c) are the results of the network deployment with 3 macro-cells, 2 pico / cell for 50 UEs/cell, and (b), (d) for 100 UEs/cell

ON-state increases with the increase in the number of users.

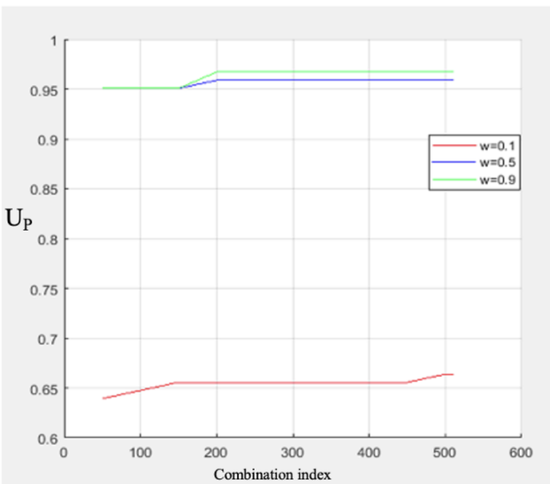
From Figs. 6.3-c and 6.3-d, we note that the number of unserved users increases at the level of combinations where the number of base stations in the off state is high, which explains that the association of users is carried out according to the availability of RBs. The high value of w $(w \geq 0.5)$ means that the reduction in power is factored than quality of service, which prove the increase in the number of unserved users with the increase of w value. Moreover, For the deployment of 50 UE / cell (Figs. 6.3 (c)), we infer that the number of unserved users is lower than that recorded in the deployment of 100 UE cell (Figs. 6.3 (d)).



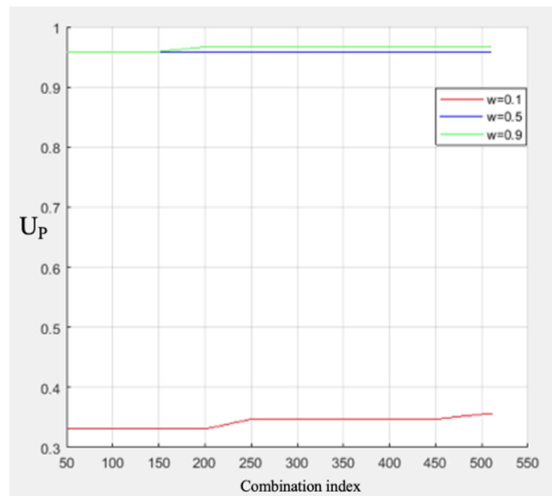
(a) U_Q **vs. w.**



(b) U_Q **vs. w.**



(c) U_P **vs. w.**

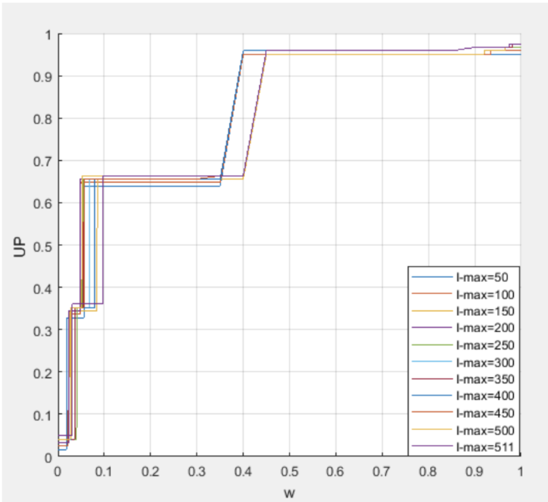


(d) U_P **vs. w.**

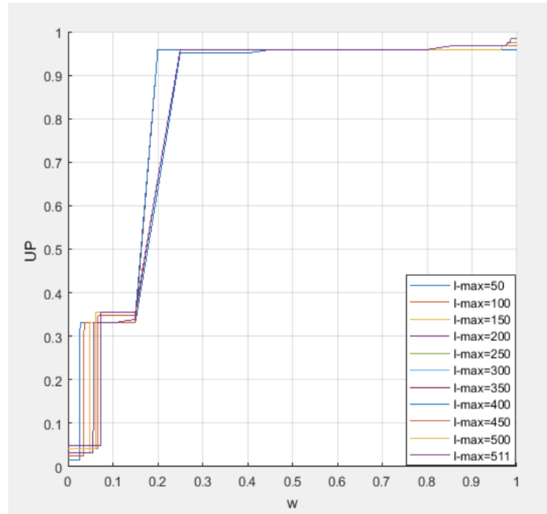
FIGURE 6.4 : (a), (c) are the results of the network deployment with 3 macro-cells, 2 pico / cell for 50 UEs/cell, and (b), (d) for100 UEs/cell

Figs. 6.4-a, 6.4-b, 6.4-c, and 6.4-d indicate the results, with respect to the gain in QoS and power, for two different configurations of users' number used. However, each combination amounts to a possible switching solution for different values of w.

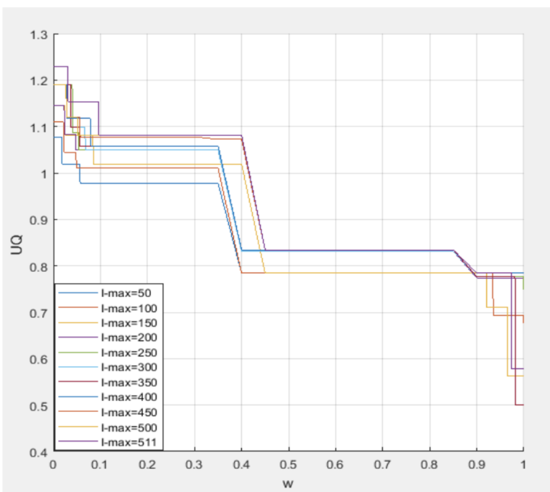
We realize that the solutions depend on different values of w. However, according to Figs. 6.5 (a), (c), we notice that the solutions retained, when $w \leq 0.5$, guarantee the QoS and even find an optimal solution which improves the throughput by 5 to 20%. Indeed, the algorithm leans towards the switching of underused base stations which induces the reduction of interference on the rest of the active base stations increasing the total QoS of the network. At the same time, the power gain is no better than that guaranteed by $w > 0.5$ since the algorithm tends to turn off more base stations, including non-used BSs, which decreases QoS.



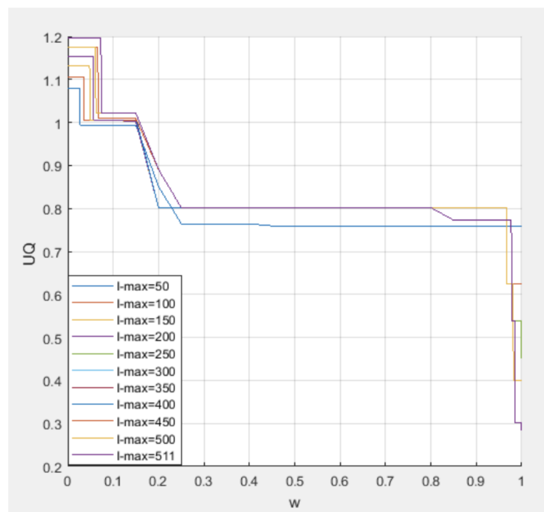
(a) U_P **as a function of several values of w.**



(b) U_P **as a function of several values of w.**



(c) U_Q **as a function of several values of w.**



(d) U_Q **as a function of several values of w.**

FIGURE 6.5 : (a), (c)are the results of the network deployment with 3 macro-cells, 2 pico / cell for 50 UEs/cell, and (b), (d)for100 UEs/cell

Let us consider a network that deploys 7 MBSs and 1 pico-BS (i.e., 16383 the total number of combinations which is equivalent to all possible commutations) as illustrated in Fig. 6.6.

For the given values of $i_{max} = [1000 : 1000 : 16383]$, Fig. 6.7-a and Fig. 6.7-b represent the utility function U_T as a function of w. I-opt indicates the optimal number of iterations in the interval $I = [1,$i_max$]$ where we obtain the maximum value V-max of the utility function.

We draw the same conclusions in the case of a deployment of 7 Macro-BS and 1 pico-BS. In Fig. 6.7 (a), the maximum value of U_T (when $w = 0.1$) is recorded with only 4 MBS in the on state, while for $w = 0.5$and for $w = 0.7$ the maximal values are recorded with 7 SBS, 6 SBS at on state, respectively, and for $w = 0.9$ with the combination of 4SBS in on state. Equally in Fig. 6.6 (b), the maximum value of U_T (when $w = 0.1$) is recorded with only 6MBS and 1 SBS in the on state, while

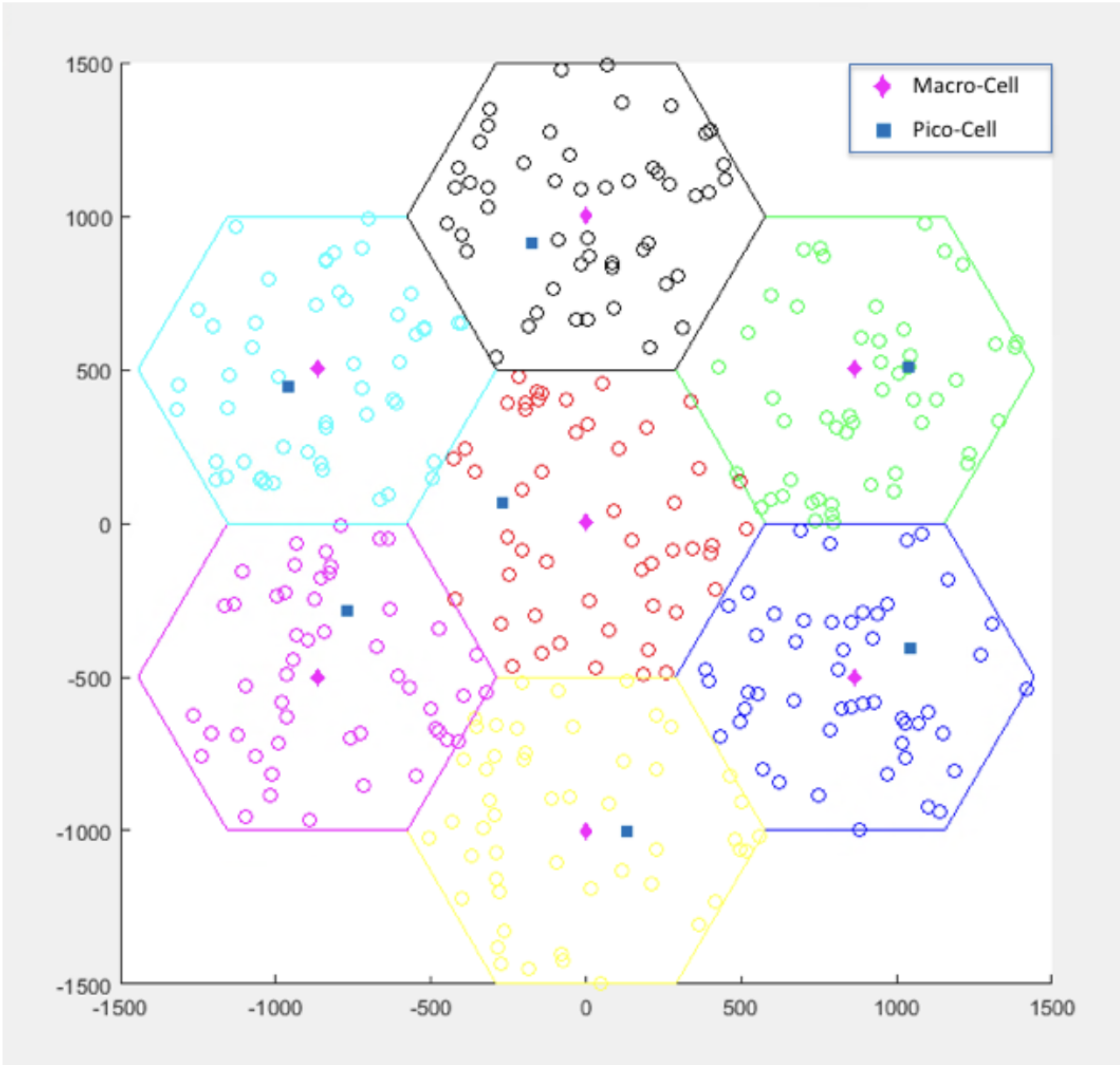
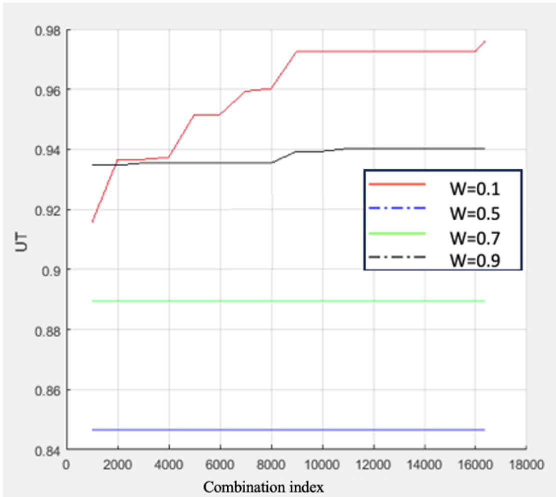


FIGURE 6.6 : Distribution of users in the cellular network with hexagonal architecture

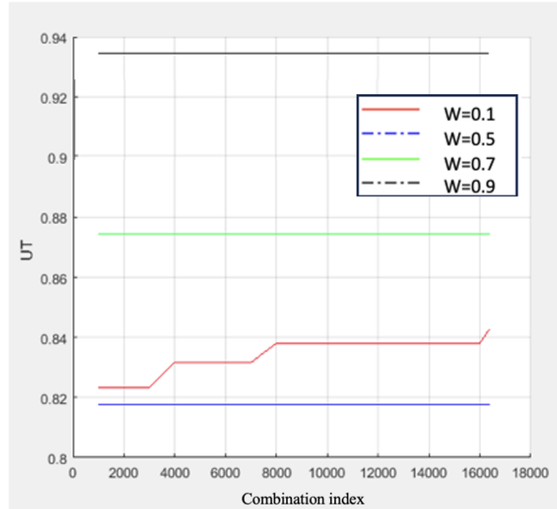
for $w = 0.5$, $w = 0.7$, and $w = 0.9$ the maximal value is recorded with 7 SBS at ON-state. Thus, it can be concluded that the number of BS in ON-state increases with the increase in the number of users.

In order to simplify our exhaustive research, we have observed that the curves of the QoS gains (as indicated in Figs. 6.8 (a), (b)) have a cyclical shape. A study of the highest values of U_Q (As shown below) was made to figure out if they have any common characteristics.

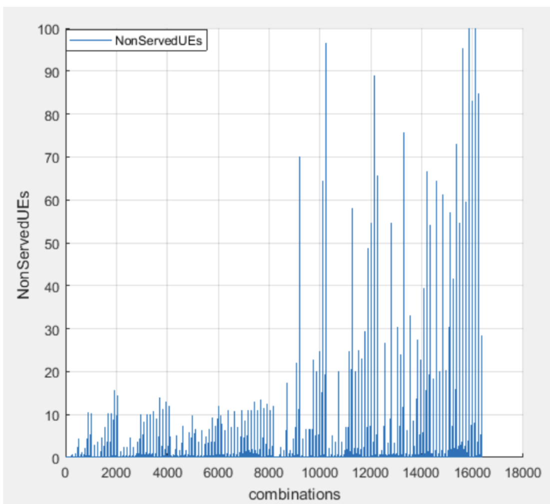
This study confirmed that all high values of U_Q are recorded when all Macro base stations are in the on state. With this analysis, we can reduce our search process by keeping all macro BSs (MBS) in ON-state and activating the process of switching through the pico BSs (SBS).



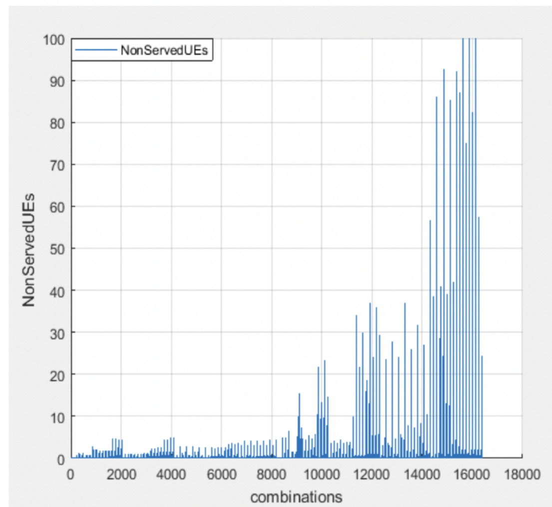
(a) Utility function(5.7), for several values of w.



(b) Utility function(5.7), for several values of w.



(c) Non served UEs.

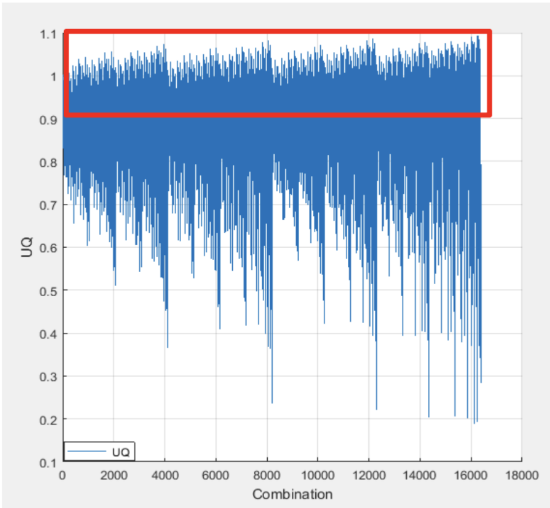


(d) Non served UEs.

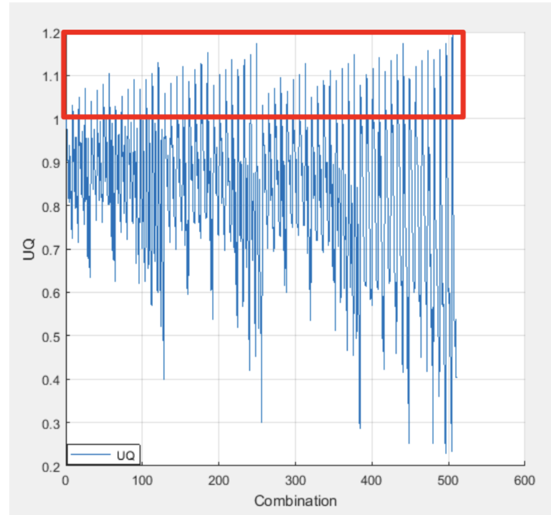
FIGURE 6.7 : (a), (c) are the results of the network deployment with 7 macro-cells, 1 pico / cell for 50 UEs/cell and (b), (d), for100 UEs/cell.

keeping all MBS active, will maintain a general coverage to the network. Therefore, it will bring better performance and ensure good quality of service.

Brute force method involves exhaustive computation and evaluation of every possible solution, which becomes impractical as the problem's complexity increases, particularly in scenarios like optimizing base station configurations where the number of variables multiplies exponentially. In contrast, the proposed DL algorithm introduce a data-driven approach. By using historical data and patterns, it can learn to recognize underlying relationships and make predictions based on this acquired knowledge. In the context of base station switching optimization, the deep learning (DL) algorithm uses a dataset that incorporates various parameters and environmental factors to intelligently predict optimal configurations, greatly reducing the computational burden while achieving



(a) U_Q with 7 MBS / 1SBS..



(b) U_Q with 3 MBS / 2SBS.

FIGURE 6.8 : U_Q per combination for the network.

better results.

6.2.2 Deep Learning Approaches Results

6.2.2.1 Related Work

Inspired by the novel solution proposed in the related work (73), which aims to enhance the energy efficiency of wireless networks through machine learning, our research undertakes a similar exploration. The authors of the benchmarked work have trained neural networks to forecast which base stations can enter a sleeping mode and predict user-base station associations. The primary objective of this approach is to reduce energy consumption. In the first neural network, the input is the normalized bit rate between the users and the base stations (BS). This information is exchanged between the mobile phone and the BS before pairing. For the second neural network, the input comprises two parts. The first part is the normalized bit rate between the users and the BS. The second part is the output of the first neural network, which provides information about which BS should be turned on and which ones should be turned off. However, The limitations of this work used as a benchmark include potential constraints in addressing real-world scenarios, variations in network conditions, and scalability concerns. Additionally, the generalization of the proposed solution to diverse network architectures and configurations may be limited.

The limitations of the related work include :

Accuracy : The proposed learning framework combines two neural networks to estimate the base station (BS) state and predict user-BS associations. While this approach provides speed, It doesn't achieve a high accuracy level when compared to a centralized algorithm.

Number of Users : The algorithm is designed to optimize performance when the number of users is between 100 and 200. When there are fewer than 100 users, only the main BS is turned on, and when there are more than 200 users, all BSs are turned on. This means that the algorithm may not be suitable for scenarios with fewer than 100 or more than 200 users.

Fixed Array Size : The input array used for training the neural networks has a fixed size, which may not accurately represent the varying number of users in real-time. However, in real-world scenarios, the number of users and base stations may vary dynamically, which could impact the performance of the model. If there are fewer users at a given moment, their input values will be set to 0, potentially affecting the accuracy of the predictions.

The same power for the macro BS and small BS have been used in their model, both the macro base station (MBS) and the small base stations (SBS) had a power usage of 40dBm. The power consumption of the base stations is stable, regardless of the fluctuation in data traffic. So, this leads to a waste of energy in the network access.

The limited scale of the network deployment (1 MBS 10 SBS) may restrict the generalization of the findings and the applicability of the proposed model to larger, more complex network configurations. Real-world wireless networks often comprise a more extensive array of base stations, covering diverse geographical areas and accommodating varying user densities and traffic patterns.

Acknowledging the limitations inherent in the benchmarked work, our research embarks on the development of new strategies aimed at addressing these constraints. Recognizing the challenges posed by real-world scenarios in wireless networks, we conduct a thorough exploration employing various deep learning (DL) models. This comprehensive testing of diverse DL architectures is undertaken with the specific goal of achieving enhanced performance in practical, real-world wireless network environments. By adapting different models, our research seeks to overcome the identified limitations and provide novel insights that contribute to the advancement of more robust and adaptable solutions for optimizing the energy efficiency of wireless networks.

Concerning the configuration of our deep learning approach, we initiated our testing by configuring the initial neural network across three different models (LSTM, GRU, and the Fused LSTM and dense layers). Our objective is to identify the optimal combination of base station statuses that enhance energy efficiency while preserving service quality. To analyse the performance, different scenarios will be compared.

In our analysis of the LSTM and GRU models, we conducted tests using the received power or Signal-to-Interference-plus-Noise Ratio (SINR) as input parameters. Notably, we observed that both parameters yielded the same results. However, in the case of the fused LSTM model, we opted for a combination of input parameters. Specifically, we utilized both the received power and the associations of User Equipment (UEs) to the corresponding Base Stations (BS) as inputs. This decision was motivated by the need to capture a more holistic view of the network dynamics, enabling the model to glean insights from multiple dimensions simultaneously and thus potentially enhancing its predictive capabilities.

The first scenario comprises 1 MBS, 2 SBS, and N users distributed randomly. The results of this part are briefly summarized in Tab. 6.2, Tab. 6.3, and Tab. 6.4. The results are categorized into two groups in Tab. 6.2 : In the first group, the number of users remains constant at 100 users per macro base station. In contrast, in the second group, the number of users varies randomly within the range of $[1, 300]$. The Tab. II presents the performance evaluation of three distinct models (LSTM, GRU, and fused LSTM) tested on a dataset comprising 10,000 samples. The table displays the training and testing accuracy values for various settings of the parameter "w" (is the power cost weight indicated in equation (5.7)). Notably, the results indicate a higher accuracy for the UE (User Equipment) variable compared to the UE fixe (Fixed User Equipment=100 UEs), signifying that the models effectively learn with variable user data. Furthermore, it is evident that the fused LSTM model consistently outperforms both the LSTM and GRU models across all values of w showcasing its superior learning capabilities.

We note that the results for $w = 0.5$ and $w = 0.9$ exhibit inferior performance compared to $w = 0.1$ when considering the throughput gain. This finding suggests that a lower value of w optimizes the throughput gain, emphasizing the importance of parameter selection in achieving desirable performance metrics.

The comparative Tab. 6.3 presents the results obtained of the three models, considering datasets of both 10,000 and 50,000 samples, with a focus on the parameter "$w = 0.1$" for the UE variable. The results underscore the substantial impact of increased data volume, with all three models displaying enhanced performance metrics when trained on the larger dataset. Specifically, the LSTM model achieves an accuracy of 94.24%, while the GRU model attains 94.95%, and the fused model remarkably outperforms both, yielding an impressive accuracy of 99.10%.

TABLEAU 6.2 : Training and testing accuracy, 10.000 samples, 800 epochs

LSTM Model				
UEs per cell	100		Variable	
w	Train	Test	Train	Test
0.1	95.81%	92.43 %	92.85%	86.90%
0.5	80.29%	66.85%	91.19%	85.68%
0.9	80.20%	66.45%	90.39%	84.53%
GRU Model				
UEs per cell	100		Variable	
w	Train	Test	Train	Test
0.1	96.75%	91.76%	93.71%	85.70%
0.5	81.54%	67.11%	92%	85.78%
0.9	82.88%	66.08%	91.30%	84.38%
Fused Model				
UEs per cell	100		Variable	
w	Train	Test	Train	Test
0.1	100%	93.43%	98.99%	88.52%
0.5	100%	68.83%	98.99%	87.38%
0.9	100%	68.53%	99.19%	87.22%

Tab. 6.4 compares the performance of the three models, trained on datasets comprising 10,000 and 50,000 samples. Notably, the focus is on the "$w = 0.5$" parameter for the UE fixe=100UEs. The consistent trend within the results showcases an evident performance improvement with larger datasets. Specifically, the LSTM model achieves an accuracy of 91.32% when trained on the 50,000-sample dataset. GRU model's results closely resemble those of the LSTM model, while the fused LSTM model outperform both LSTM and GRU model and reach approximately 100%.

These findings underscore the significant impact of larger datasets on the overall performance of the models, indicating that a higher volume of data facilitates a more comprehensive learning

TABLEAU 6.3 : Training and testing accuracies

$w = 0.1$				
Samples	10000		50000	
Model	Train	Test	Train	Test
LSTM Model	92.85%	86.90%	94.24%	94.24%
GRU Model	93.71%	85.70%	94.95%	94.95%
Fused Model	98.99%	88.52%	99.10%	99.10%

TABLEAU 6.4 : Training and testing accuracies

$w = 0.5, 100\,UEs$				
Samples	10000		50000	
Model	Train	Test	Train	Test
LSTM Model	80.29%	66.85%	91.32%	91.32%
GRU Model	81.54%	67.11%	93.92%	93.92%
Fused Model	100%	68.83%	100%	100%

process, leading to improved predictive accuracy and reliability.

In the second scenario, we adopt the same configuration (1MBS and 10 SBS with a fixed number of 150 UEs) as used in reference (73) to demonstrate the performance of our results based on various criteria.

TABLEAU 6.5 : Accuracy with 1000 samples

$w = 0.1, 150\,UEs$		
Epochs	LSTM/GRU Model	Fused model
10	81.63%	100%
800	100%	100%

Tab. 6.5 presents the accuracy results for 1000 samples across the three distinct ML models, highlighting the impact of different epoch values on performance. In the context of ML, epochs re-

present the number of times a learning algorithm undergoes training on the entire dataset. Notably, the findings reveal that while the LSTM and GRU models achieve a perfect 100% accuracy with a higher epoch value of 800, they only reach a modest 81.63% accuracy with 10 epochs. In contrast, the proposed fused model notably outperforms the model described in (73), which achieved only 93% accuracy, swiftly attaining 100% accuracy with just 10 epochs.

TABLEAU 6.6 : Accuracy with 10.000 samples

$w = 0.1, 150\,UEs$	
LSTM/GRU (50 epochs)	Fused Model (4 epochs)
100%	100%

Moreover, Tab. 6.6 confirms that with a larger dataset (10000 samples), both the LSTM and GRU models can achieve 100% accuracy with just 50 epochs, compared to 800 epochs as shown in Tab. V with a smaller dataset. In contrast, the fused model can achieve 100% accuracy with only 4 epochs.

These results underscore the superior learning capability of the fused model, highlighting its potential to achieve optimal performance in the given context. Moreover, the reduced number of epochs not only ensures superior accuracy but also effectively reduces the model's complexity, emphasizing the efficiency and practicality of the proposed fused model.

The limited scale of the network deployment may restrict the general applicability of the proposed model to larger, more complex network configurations. Real-world wireless networks often comprise a more extensive array of base stations, covering diverse geographical areas and accommodating varying user densities and traffic patterns. By using a relatively small-scale network setup, the study might not fully capture the challenges and dynamics associated with managing energy efficiency in more extensive and heterogeneous network environments. Additionally, the specific characteristics of a network with only one macro and 10 small base stations may not adequately represent the complexities and intricacies of practical network deployments, potentially limiting the model's ability to handle diverse network topologies and traffic variations effectively.

To enhance the robustness and reliability of the proposed approach, we explore the implications of the model when applied to larger and more diverse network configurations. Scaling up the study to encompass a more comprehensive range of base station types, network sizes, and geographical distributions would provide a more comprehensive understanding of the model's performance under varying network conditions and further validate its effectiveness in optimizing energy efficiency across different deployment scenarios.

TABLEAU 6.7 : Trainning and testing accuracy, 30.000 samples, 800 epochs

LSTM Model				
UEs per cell	100		Variable	
w	Train accuracy	Test accuracy	Train accuracy	Test accuracy
0.1	100%	100%	99.51%	99.51%
0.5	100%	100%	99.56%	99.56%
0.9	100%	100%	99.57%	99.57%
GRU Model				
UEs per cell	100		Variable	
w	Train accuracy	Test accuracy	Train accuracy	Test accuracy
0.1	100%	100%	99.56%	99.56%
0.5	100%	100%	99.47%	99.47%
0.9	100%	100%	99.62%	99.62%
Fused Model				
UEs per cell	100		Variable	
w	Train accuracy	Test accuracy	Train accuracy	Test accuracy
0.1	100%	100%	99.92%	99.92%
0.5	100%	100%	99.92%	99.92%
0.9	100%	100%	99.92%	99.92%

In the third scenarios, a network with a hexagonal architecture is tested, comprising three Macro Base Stations (MBSs) and two Small Base Stations (SBSs) per cell, with 'N' UEs per MBS. The MBS is placed in the center of the hexagon, whereas the SBSs are randomly positioned.

The comparison between the findings in Tab. 6.2 and Tab. 6.7 indicates that the configuration involving the three MBSs forming 3 cells within a heterogeneous network yields promising results. Notably, the network achieves a remarkable 100% accuracy for UE fixe and demonstrates a high accuracy rate of 99.92% for UE variable. These outcomes suggest the effectiveness of the hexagonal network configuration in ensuring excellent performance in terms of accuracy, highlighting the network's capability to manage and accommodate different types of scenarios within the system.

The comprehensive comparison within these results demonstrates the distinct advantages of the fused LSTM model over the LSTM and GRU models, affirming its potential for enhancing the learning process in this context.

In summary, the fused model's functionality provides flexibility in modeling and predicting BSs states. The network can adapt to various scenarios and learn from multiple modalities, leading to more precise predictions. Overall, the fused LSTM and Dense layers, with LeakyReLU activation functions, offer improved learning capabilities, nonlinear transformations, the utilization of complementary information, and enhanced flexibility in modeling and prediction.

TABLEAU 6.8 : NMSE for throughput prediction in different Scenarios

$w = 0.1$		
UEs per cell	Variable	100
LSTM	0.1205	0.0110
GRU	0.1026	0.0110
Fused Model	0.0414	0.0022

In the first scenario, involving 1 Macro BS and 2 Pico cells with a weight factor w of 0.1 and a random number of users, the LSTM model achieved an NMSE of 0.1205, indicating a moderate level of error in predicting throughput based on received power. The GRU model performed slightly better with an NMSE of 0.1026, while the fused model demonstrated a significant improvement with an NMSE of 0.0414. In the second scenario, where the number of users was fixed at 100, the LSTM and GRU models both exhibited exceptional accuracy, with NMSE values of 0.0110 each. The fused model continued to enhance predictive performance, yielding an even lower NMSE of 0.0022. These results from Tab. 6.8 indicate that, in scenarios with a fixed number of users, the fused model consistently outperforms individual LSTM and GRU models, showcasing its effectiveness in improving predictive accuracy for throughput prediction based on received power or SINR.

6.3 Conclusion

In our study, we initially employed the brute force algorithm to identify the optimal combination of on/off base stations. This algorithm maximizes the utility function by striking a balance between the gains in throughput and the reductions in power consumption. The brute force approach served as a benchmark for our subsequent work involving deep learning techniques. Within our deep learning model, we utilized three distinct architectures : Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a fused model (combining LSTM and dense layers). Through extensive testing across various scenarios, we applied LSTM and GRU to predict the optimal configuration of active base stations based on received power levels as input parameter. Notably, we observed that increasing the number of epochs or expanding the training dataset enhanced the prediction accuracy for both LSTM and GRU. The 'fused' model, incorporating an additional parameter as input representing user associations with base stations, achieved a remarkable 100% prediction accuracy. This result signifies the attainment of an optimal combination that maximizes gains in both throughput and power, aligning with the outcomes generated by the brute force algorithm.

Consequently, our work empowers us to achieve energy efficiency levels comparable to those of the BF algorithm, while significantly reducing computational costs in terms of both time and resources. A prominent advantage of this model lies in its ability to save time, aligning with one of the primary objectives of our research : adapting optimization outcomes to real-world environments, enabling swift adjustments of mobile cells in response to dynamic user changes.

7 Conclusion and Future Work

7.1 Conclusions

To face the exponential increase of traffic demands from cellular networks and the subsequent potential danger of more damage to our environment due to the release of substantial amounts of greenhouse gases, the reduction of OPEX by improving the energy efficiency in such networks will be extremely important for service providers. So, from an operator's perspective, reducing energy consumption will also translate to lower operating expenditure (OPEX) costs, which will lead to reducing carbon emissions for wireless cellular networks. In fact, to obtain significant reductions in CO2 emissions in the upcoming years, the Green Radio program is one of the most important aspirations of achieving a hundredfold reduction in power consumption over current designs for wireless communication networks.

In this context, our work represents a substantial advancement in the realm of energy efficiency in wireless communications. Our approach achieves efficiency levels comparable to the well-established Brute Force (BF) algorithm while significantly reducing computational costs in terms of time and resources. The introduction of Deep Multimodal Learning (DML) in wireless communications marks a novel endeavor to enhance energy efficiency. Simulation results underscore the superior performance of DML-based architectures over conventional approaches, emphasizing the potential of leveraging multi-modal data.

A notable advantage of our proposed model lies in its time-saving capability, aligning seamlessly with one of our primary research objectives : adapting optimization outcomes to real-world environments. This adaptability enables swift adjustments of mobile cells in response to dynamic user changes, a crucial feature for the evolving landscape of wireless communication.

Looking forward, our proposed Base Station (BS) on/off switching algorithms hold particular promise for future-generation networks characterized by extreme BS densification, introducing a new layer of complexity to the energy efficiency challenge. The suggested approaches stand out for their flexibility, adaptability, low complexity, and scalability. They can effortlessly incorporate additional optimization criteria (i.e., CAPEX & OPEX costs, etc.) by simply adjusting the proposed metric.

In conclusion, our research not only addresses the pressing need for energy-efficient wireless communication networks but also paves the way for innovative solutions that balance environmental concerns with the evolving demands of modern communication systems.

7.2 Future Work Directions

The efforts undertaken in this thesis within various aspects of environmentally friendly cellular networks represent progress towards aligning 5G networks to meet the rigorous end-user application demands while ensuring a minimal energy impact. This objective encompasses several goals, including the reduction of carbon emissions, grid energy consumption, and operational electric costs for network operators. Our research has revealed numerous unexplored opportunities within this domain as per bellow :

7.2.1 Enhance the BS Switch-Off Strategy in HetNets by Effectively Utilizing a Combination of Diverse Criteria

As a future work, our focus will be on further exploring the implementation of the BS switch-off strategy as a means to enhance energy efficiency and performance in HetNets. To ensure the development of more sustainable and resilient 5G systems, we plan to investigate the introduction of additional measures that leverage the combination of various criteria, including resource allocation, inter-site carrier aggregation (CA), and load-aware strategies. Additionally, our research will delve into the design and examination of strategies that effectively utilize a fusion of different criteria and perspectives, such as the combination of load and distance criteria , or random and load criteria , among others.

7.2.1.1 Incorporate the Quality of Experience (QoE) into the System Design

The imminent widespread deployment of 5G networks calls for efficient management of control overhead and better utilization of shared bandwidth. Integrating enhanced Quality of Service (QoS) constraints is crucial for meeting user satisfaction, considering various QoS factors including outage probability, download rate, congestion probability, delay, video type and format, and energy consumption of mobile devices. To address this, the proposed work will focus on introducing wireless access virtualization and a novel scheduling approach to prioritize individual user's QoS requirements and dynamic adaptive clustering. This approach aims to create user-specific virtual base stations (uVBSs) that cater to each user's environment and QoS needs, ensuring balanced traffic distribution and seamless communication without cell-edge effects. The objective is to establish effective resource allocation strategies to achieve fairness, maximum spectral efficiency, and QoS.

7.2.2 Renewable Energy Sources for Powering Base Stations

To optimize the performance of the BS switched on/off approach and mitigate the environmental impact of mobile networks, particularly in terms of energy consumption and cost savings, a novel solution integrating renewable energies with the electrical grid for powering wireless systems is proposed. Nevertheless, this strategy encounters uncertainties associated with renewable power generation, power pricing, and wireless traffic load. To address these challenges, an adaptive demand-side power management scheme will be developed, enabling intelligent decision-making on energy consumption between renewable sources and grid energy. The aim is to achieve highly efficient energy utilization while simultaneously reducing operating expenses and greenhouse gas emissions.

7.2.3 Machine Learning and Data Correlations

Given the advantages of machine learning in enhancing performance and simplifying complex problem-solving, it becomes evident that it offers a feasible alternative to conventional algorithmic approaches. The ability to learn from the environment stands as its primary advantage. However, the scarcity of research datasets and the challenges associated with data acquisition from networks remain pressing issues.Preparing the model for training entails precise data alignment, debugging, and the removal of any biased values, necessitating significant time and effort. Future research should prioritize the trade-off between efficient machine learning for wireless networks and the simplification of models, particularly in regions where energy efficiency is of paramount significance.

In conclusion, the investigation into sleep mode technologies and the broader realm of green cellular networks continues to be an active and essential field of research. Given the growing emphasis on energy efficiency in the present era, these research domains are expected to remain relevant, presenting substantial prospects for advancements and the resolution of pertinent challenges.

BIBLIOGRAPHIE

[1] Abuibaid MA, Çolak SA (2017) Energy-efficient massive mimo system : Exploiting user location distribution variation. *AEU-International Journal of Electronics and Communications*, 72:17–25.

[2] Afshang M, Dhillon HS (2018) Poisson cluster process based analysis of hetnets with correlated user and base station locations. *IEEE Transactions on Wireless Communications*, 17(4):2417–2431.

[3] Aggarwal A, Selvamuthu D (2023) Energy optimization of a base station using q-learning algorithm. *2023 17th International Conference on Telecommunications (ConTEL)*, IEEE, pages 1–5.

[4] Ahmed E, Rehmani MH (2017) *Mobile edge computing : opportunities, solutions, and challenges*.

[5] Akbar S, Deng Y, Nallanathan A, Elkashlan M, Aghvami AH (2016) Simultaneous wireless information and power transfer in k-tier heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, 15(8):5804–5818.

[6] Alamu O, Gbenga-Ilori A, Adelabu M, Imoize A, Ladipo O (2020) Energy efficiency techniques in ultra-dense wireless heterogeneous networks : An overview and outlook. *Engineering Science and Technology, an International Journal*, 23(6):1308–1326.

[7] Alhumaima R (2017) *Evaluation of power consumption and trade-offs in 5G mobile communications networks*. Thèse de doctorat, Brunel University London.

[8] Alhumaima RS, Al-Raweshidy HS (2016) Evaluating the energy efficiency of software defined-based cloud radio access networks. *IET Communications*, 10(8):987–994.

[9] Ali S, Saad W, Rajatheva N, Chang K, Steinbach D, Sliwa B, Wietfeld C, Mei K, Shiri H, Zepernick HJ et al. (2020) 6g white paper on machine learning in wireless communication networks. *arXiv preprint arXiv :2004.13875*.

[10] Alliance N (2015) 5g white paper. *Next generation mobile networks, white paper*, 1(2015).

[11] Alsharif MH, Kelechi AH, Albreem MA, Chaudhry SA, Zia MS, Kim S (2020) Sixth generation (6g) wireless networks : Vision, research activities, challenges and potential solutions. *Symmetry*, 12(4):676.

[12] Alsharif MH, Nordin R, Ismail M (2014) Classification, recent advances and research challenges in energy efficient cellular networks. *Wireless Personal Communications*, 77:1249–1269.

[13] An L, Zhang T, Feng C (2015) Stochastic geometry based energy-efficient base station density optimization in cellular networks. *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, pages 1614–1619.

[14] Arani AH, Omidi MJ, Mehbodniya A, Adachi F (2017) Minimizing base stations' on/off switchings in self-organizing heterogeneous networks : A distributed satisfactory framework. *IEEE Access*, 5:26267–26278.

[15] Arnold O, Richter F, Fettweis G, Blume O (2010) Power consumption modeling of different base station types in heterogeneous cellular networks. *2010 Future Network & Mobile Summit*, IEEE, pages 1–8.

[16] Arvaje FR, Ghahfarokhi BS (2017) A spectrum efficient base station switching-off mechanism for green cellular networks. *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, pages 433–438.

[17] Ashmore R, Calinescu R, Paterson C (2021) Assuring the machine learning lifecycle : Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5):1–39.

[18] Auer G, Blume O, Giannini V, Godor I, Imran M, Jading Y, Katranaras E, Olsson M, Sabella D, Skillermark P et al. (2010) Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. earth proj. *D2*, 3:39–40.

[19] Auer G, Giannini V, Desset C, Godor I, Skillermark P, Olsson M, Imran MA, Sabella D, Gonzalez MJ, Blume O et al. (2011) How much energy is needed to run a wireless network? *IEEE wireless communications*, 18(5):40–49.

[20] Azari A, Papapetrou P, Denic S, Peters G (2019) Cellular traffic prediction and classification : A comparative evaluation of lstm and arima. *Discovery Science : 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, Springer, pages 129–144.

[21] Bacci G, Belmega EV, Mertikopoulos P, Sanguinetti L (2015) Energy-aware competitive power allocation for heterogeneous networks under qos constraints. *IEEE Transactions on Wireless Communications*, 14(9):4728–4742.

[22] Badic B, O'farrrell T, Loskot P, He J (2009) Energy efficient radio access architectures for green radio : Large versus small cell size deployment. *2009 IEEE 70th vehicular technology conference fall*, IEEE, pages 1–5.

[23] Banafaa M, Shayea I, Din J, Azmi MH, Alashbi A, Daradkeh YI, Alhammadi A (2023) 6g mobile communication technology : Requirements, targets, applications, challenges, advantages, and opportunities. *Alexandria Engineering Journal*, 64:245–274.

[24] Björnson E, Sanguinetti L, Hoydis J, Debbah M (2015) Optimal design of energy-efficient multi-user mimo systems : Is massive mimo the answer? *IEEE Transactions on wireless communications*, 14(6):3059–3075.

[25] Björnson E, Sanguinetti L, Kountouris M (2016) Deploying dense networks for maximal energy efficiency : Small cells meet massive mimo. *IEEE Journal on Selected Areas in Communications*, 34(4):832–847.

[26] Blume O, Eckhardt H, Klein S, Kuehn E, Wajda WM (2010) Energy savings in mobile networks based on adaptation to traffic statistics. *Bell Labs Technical Journal*, 15(2):77–94.

[27] Bousia A, Antonopoulos A, Alonso L, Verikoukis C (2012) " green" distance-aware base station sleeping algorithm in lte-advanced. *2012 IEEE international conference on communications (ICC)*, IEEE, pages 1347–1351.

[28] Busari SA, Huq KMS, Mumtaz S, Dai L, Rodriguez J (2017) Millimeter-wave massive mimo communication for future wireless systems : A survey. *IEEE Communications Surveys & Tutorials*, 20(2):836–869.

[29] Buzzi S, Colavolpe G, Saturnino D, Zappone A (2011) Potential games for energy-efficient power control and subcarrier allocation in uplink multicell ofdma systems. *IEEE Journal of Selected Topics in Signal Processing*, 6(2):89–103.

[30] Chakareski J, Naqvi S, Mastronarde N, Xu J, Afghah F, Razi A (2019) An energy efficient framework for uav-assisted millimeter wave 5g heterogeneous cellular networks. *IEEE Transactions on Green Communications and Networking*, 3(1):37–44.

[31] Chen X, Jiao L, Li W, Fu X (2015a) Efficient multi-user computation offloading for mobile-edge cloud computing. *IEEE/ACM transactions on networking*, 24(5):2795–2808.

[32] Chen X, Wu J, Cai Y, Zhang H, Chen T (2015b) Energy-efficiency oriented traffic offloading in wireless networks : A brief survey and a learning approach for heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(4):627–640.

[33] Chen X, Wu X, Han S, Xie Z (2019) Joint optimization of ee and se considering interference threshold in ultra-dense networks. *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, IEEE, pages 1305–1310.

[34] Cheung KTK, Yang S, Hanzo L (2013) Achieving maximum energy-efficiency in multi-relay ofdma cellular networks : A fractional programming approach. *IEEE Transactions on Communications*, 61(7):2746–2757.

[35] Chia YK, Sun S, Zhang R (2014) Energy cooperation in cellular networks with renewable powered base stations. *IEEE Transactions on Wireless Communications*, 13(12):6996–7010.

[36] Chiaraviglio L, Ciullo D, Meo M, Marsan MA, Torino I (2008) *Energy-aware UMTS access networks*.

[37] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*.

[38] Cho KJ, Kim JH, Stapleton SP (2005) A highly efficient doherty feedforward linear power amplifier for w-cdma base-station applications. *IEEE Transactions on Microwave Theory and Techniques*, 53(1):292–300.

[Choras] Choras M. 5g communications : Energy efficiency. available :https ://downloads.hindawi.com/journals/specialissues/478121.pdf.

[40] Correia LM, Zeller D, Blume O, Ferling D, Jading Y, Gódor I, Auer G, Van Der Perre L (2010) Challenges and enabling technologies for energy aware mobile radio networks. *IEEE Communications Magazine*, 48(11):66–72.

[41] Dai B, Yu W (2016) Energy efficiency of downlink transmission strategies for cloud radio access networks. *IEEE Journal on Selected Areas in Communications*, 34(4):1037–1050.

[42] De Domenico A, Kténas D (2018) Reinforcement learning for interference-aware cell dtx in heterogeneous networks. *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, pages 1–6.

[43] Debaillie B, Desset C, Louagie F (2015) A flexible and future-proof power model for cellular base stations. *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, IEEE, pages 1–7.

[44] Demirtaş M, Soysal A et al. (2017) Nonoverlay heterogeneous network planning for energy efficiency. *Wireless Communications and Mobile Computing*, 2017.

[45] Ding H, Zhang H, Tian J, Xu S, Yuan D (2018) Energy efficient user association and power control for dense heterogeneous networks. *2018 International Conference on Computing, Networking and Communications (ICNC)*, IEEE, pages 741–746.

[46] Dong X, Zheng FC, Zhu X, Luo J (2019) Hetnets with range expansion : Local delay and energy efficiency optimization. *IEEE Transactions on Vehicular Technology*, 68(6):6147–6150.

[47] Du B, Pan C, Zhang W, Chen M (2014) Distributed energy-efficient power optimization for comp systems with max-min fairness. *IEEE Communications Letters*, 18(6):999–1002.

[48] Edler T (2008) Green base stations—how to minimize co2 emission in operator networks. *Ericsson seminar, Bath Base Station Conf*.

[49] Ekti AR, Shakir MZ, Serpedin E, Qaraqe KA (2013) Downlink power consumption of hetnets based on the probabilistic traffic model of mobile users. *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, pages 2797–2802.

[50] El-Amine A, Hassan HAH, Iturralde M, Nuaymi L (2019) Location-aware sleep strategy for energy-delay tradeoffs in 5g with reinforcement learning. *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, IEEE, pages 1–6.

[51] El Morabit Y, Mrabti F, Abarkan EH (2017) Small cell switch off using genetic algorithm. *2017 international conference on advanced technologies for signal and image processing (ATSIP)*, IEEE, pages 1–4.

[52] Fang F, Cheng J, Ding Z (2018a) Joint energy efficient subchannel and power optimization for a downlink noma heterogeneous network. *IEEE Transactions on Vehicular Technology*, 68(2): 1351–1364.

[53] Fang Z, Wu Y, Lu Y, Hu J, Peng T, Ye J (2018b) Simultaneous wireless information and power transfer in cellular two-way relay networks with massive mimo. *IEEE Access*, 6:29262–29270.

[54] Fehske AJ, Marsch P, Fettweis GP (2010) Bit per joule efficiency of cooperating base stations in cellular networks. *2010 IEEE Globecom workshops*, IEEE, pages 1406–1411.

[55] Fehske AJ, Richter F, Fettweis GP (2009) Energy efficiency improvements through micro sites in cellular mobile radio networks. *2009 IEEE Globecom Workshops*, IEEE, pages 1–5.

[56] Feng D, Jiang C, Lim G, Cimini LJ, Feng G, Li GY (2012) A survey of energy-efficient wireless communications. *IEEE Communications Surveys & Tutorials*, 15(1):167–178.

[57] Feng M, Mao S, Jiang T (2017) Boost : Base station on-off switching strategy for green massive mimo hetnets. *IEEE Transactions on Wireless Communications*, 16(11):7319–7332.

[58] Fowdur TP, Doorgakant B (2023) A review of machine learning techniques for enhanced energy efficient 5g and 6g communications. *Engineering Applications of Artificial Intelligence*, 122:106032.

[59] Fu R, Zhang Z, Li L (2016) Using lstm and gru neural network methods for traffic flow prediction. *2016 31st Youth academic annual conference of Chinese association of automation (YAC)*, IEEE, pages 324–328.

[60] Fu Y, Wang S, Wang CX, Hong X, McLaughlin S (2018) Artificial intelligence to manage network traffic of 5g wireless networks. *IEEE network*, 32(6):58–64.

[61] Gan X, Wang L, Feng X, Liu J, Yu H, Zhang Z, Liu H (2015) Energy efficient switch policy for small cells. *China Communications*, 12(1):78–88.

[62] Gao D, Liang Z, Zhang H, Dobre OA, Karagiannidis GK (2018) Stackelberg game-based energy efficient power allocation for heterogeneous noma networks. *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pages 1–5.

[63] Gao J, Ren Q, Gu PS, Song X (2019) User association and small-cell base station on/off strategies for energy efficiency of ultradense networks. *Mobile Information Systems*, 2019.

[64] Gao X, Dai L, Han S, Chih-Lin I, Heath RW (2016) Energy-efficient hybrid analog and digital precoding for mmwave mimo systems with large antenna arrays. *IEEE Journal on Selected Areas in Communications*, 34(4):998–1009.

[65] Ge X, Yang J, Gharavi H, Sun Y (2017) Energy efficiency challenges of 5g small cell networks. *IEEE communications Magazine*, 55(5):184–191.

[66] Gelenbe E, Caseau Y (2015) The impact of information technology on energy consumption and carbon emissions. *ubiquity*, 2015(June):1–15.

[67] Ghazzai H, Yaacoub E, Alouini MS, Abu-Dayya A (2014) Optimized smart grid energy procurement for lte networks using evolutionary algorithms. *IEEE Transactions on vehicular technology*, 63(9):4508–4519.

[68] Ghazzai H, Yaacoub E, Alouini MS, Dawy Z, Abu-Dayya A (2015) Optimized lte cell planning with varying spatial and temporal user densities. *IEEE Transactions on vehicular technology*, 65(3):1575–1589.

[69] Giambene G, Bourgeau T, Chaouchi H et al. (2016) Iterative multi-level soft frequency reuse with load balancing for heterogeneous lte-a systems. *IEEE Transactions on Wireless Communications*, 16(2):924–938.

[70] Gong J, Zhou S, Niu Z, Yang P (2010) Traffic-aware base station sleeping in dense cellular networks. *2010 IEEE 18th International Workshop on Quality of Service (IWQoS)*, IEEE, pages 1–2.

[71] González González D, Mutafungwa E, Haile B, Hämäläinen J, Poveda H et al. (2017) A planning and optimization framework for ultra dense cellular deployments. *Mobile Information Systems*, 2017.

[72] Gruber N, Jockisch A (2020) Are gru cells more specific and lstm cells more sensitive in motive classification of text? *Frontiers in artificial intelligence*, 3:40.

[73] Guerra I, Yin B, Zhang S, Cheng Y (2021) Optimization of base station on-off switching with a machine learning approach. *ICC 2021-IEEE International Conference on Communications*, IEEE, pages 1–6.

[74] Guo X, Zhou S, Niu Z, Kumar P (2013) Optimal wake-up mechanism for single base station with sleep mode. *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, IEEE, pages 1–8.

[75] Han C, Harrold T, Armour S, Krikidis I, Videv S, Grant PM, Haas H, Thompson JS, Ku I, Wang CX et al. (2011) Green radio : radio techniques to enable energy-efficient wireless networks. *IEEE communications magazine*, 49(6):46–54.

[76] Han F, Zhao S, Zhang L, Wu J (2016) Survey of strategies for switching off base stations in heterogeneous networks for greener 5g systems. *IEEE Access*, 4:4959–4973.

[77] Han S, Bian S et al. (2020) Energy-efficient 5g for a greener future. *Nature Electronics*, 3(4):182–184.

[78] Hao Y, Chen M, Hu L, Hossain MS, Ghoneim A (2018) Energy efficient task caching and offloading for mobile edge computing. *Ieee access*, 6:11365–11373.

[79] He C, Li GY, Zheng FC, You X (2013a) Energy-efficient resource allocation in ofdm systems with distributed antennas. *IEEE Transactions on Vehicular Technology*, 63(3):1223–1231.

[80] He S, Huang Y, Yang L, Ottersten B (2013b) Coordinated multicell multiuser precoding for maximizing weighted sum energy efficiency. *IEEE Transactions on Signal Processing*, 62(3): 741–751.

[81] Héliot F, Imran MA, Tafazolli R (2012) On the energy efficiency-spectral efficiency trade-off over the mimo rayleigh fading channel. *IEEE Transactions on Communications*, 60(5):1345–1356.

[82] Holtkamp H, Auer G, Bazzi S, Haas H (2013) Minimizing base station power consumption. *IEEE Journal on Selected Areas in Communications*, 32(2):297–306.

[83] Huang X, Tang S, Zheng Q, Zhang D, Chen Q (2018) Dynamic femtocell gnb on/off strategies and seamless dual connectivity in 5g heterogeneous cellular networks. *IEEE Access*, 6:21359–21368.

[84] Huawei Technologies, Co. L (2020) “green 5g : Building a sustainable world,”. available : https ://www.huawei.com/en/public-policy/green-5g-building-a-sustainable-world. *Tech. Rep., Aug. 2020. [Online]*.

[85] Huo L, Jiang D, Lv Z (2018) Soft frequency reuse-based optimization algorithm for energy efficiency of multi-cell networks. *Computers & Electrical Engineering*, 66:316–331.

[86] Huo Y, Dong X, Xu W (2017) 5g cellular user equipment : From theory to practical hardware design. *IEEE Access*, 5:13992–14010.

[87] ICT (2012) “sustainable ict in corporate organizations,” available : https ://www.itu.int/dms. *Tech. Rep., 2012. [Online]*.

[88] Imran A, Zoha A, Abu-Dayya A (2014) Challenges in 5g : how to empower son with big data for enabling 5g. *IEEE network*, 28(6):27–33.

[89] ITU-T (2008) “summary of smart 2020 report,”. available : https ://www.itu.int/md/t05-fg.ict-c-0004/en. *Tech. Rep., Aug. 2008. [Online]*.

[90] Jahid A, Islam MS, Hossain MS, Hossain ME, Monju MKH, Hossain MF (2019) Toward energy efficiency aware renewable energy management in green cellular networks with joint coordination. *IEEE Access*, 7:75782–75797.

[91] Jahid A, Shams AB, Hossain MF (2018) Green energy driven cellular networks with jt comp technique. *Physical Communication*, 28:58–68.

[92] Jahid A, Shams AB, Hossain MF et al. (2017) Pv-powered comp-based green cellular networks with a standby grid supply. *International Journal of Photoenergy*, 2017.

[93] Jang S, Lee H, Kang S, Oh T, Lee I (2018) Energy efficient swipt systems in multi-cell miso networks. *IEEE Transactions on Wireless Communications*, 17(12):8180–8194.

[94] Jia Q, Xie R, Huang T, Liu J, Liu Y (2017) Energy-efficient cooperative coded caching for heterogeneous small cell networks. *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, pages 468–473.

[95] Jiang C, Zhang H, Ren Y, Han Z, Chen KC, Hanzo L (2016) Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, 24(2):98–105.

[96] Jiang X, Zheng FC (2018) User rate and energy efficiency of hetnets based on poisson cluster process. *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, IEEE, pages 1–5.

[97] Kamel M, Hamouda W, Youssef A (2016) Ultra-dense networks : A survey. *IEEE Communications surveys & tutorials*, 18(4):2522–2545.

[98] Kanwal K, Safdar GA, Ur-Rehman M, Yang X (2017) Energy management in lte networks. *IEEE Access*, 5:4264–4284.

[99] Khiadani N (2020) Vision, requirements and challenges of sixth generation (6g) networks. *2020 6th Iranian conference on signal processing and intelligent systems (ICSPIS)*, IEEE, pages 1–4.

[100] Khirallah C, Vukobratovic D, Thompson J (2012) Performance analysis and energy efficiency of random network coding in lte-advanced. *IEEE Transactions on Wireless Communications*, 11(12):4275–4285.

[101] Klessig H, Öhmann D, Reppas AI, Hatzikirou H, Abedi M, Simsek M, Fettweis GP (2016) From immune cells to self-organizing ultra-dense small cell networks. *IEEE Journal on Selected Areas in Communications*, 34(4):800–811.

[102] Lai WS, Chang TH, Lee TS (2016) Joint power and admission control for spectral and energy efficiency maximization in heterogeneous ofdma networks. *IEEE Transactions on Wireless Communications*, 15(5):3531–3547.

[103] Larsson EG, Edfors O, Tufvesson F, Marzetta TL (2014) Massive mimo for next generation wireless systems. *IEEE communications magazine*, 52(2):186–195.

[104] Lei J, Chen H, Zhao F (2018) Stochastic geometry analysis of downlink spectral and energy efficiency in ultradense heterogeneous cellular networks. *Mobile Information Systems*, 2018.

[105] Letaief KB, Chen W, Shi Y, Zhang J, Zhang YJA (2019) The roadmap to 6g : Ai empowered wireless networks. *IEEE communications magazine*, 57(8):84–90.

[106] Li L, Peng M, Yang C, Wu Y (2015) Optimization of base-station density for high energy-efficient cellular networks with sleeping strategies. *IEEE Transactions on Vehicular Technology*, 65(9):7501–7514.

[107] Li L, Zhao G, Blum RS (2018) A survey of caching techniques in cellular networks : Research issues and challenges in content placement and delivery strategies. *IEEE Communications Surveys & Tutorials*, 20(3):1710–1732.

[108] Li M, Wang Z, Li H, Liu Q, Zhou L (2019) A hardware-efficient hybrid beamforming solution for mmwave mimo systems. *IEEE Wireless Communications*, 26(1):137–143.

[109] Li Z, Yu P, Li W, Qiu X (2016) Modeling and optimization of self-organizing energy-saving mechanism for hetnets. *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*, IEEE, pages 146–152.

[110] Lin Z, Du X, Chen HH, Ai B, Chen Z, Wu D (2019) Millimeter-wave propagation modeling and measurements for 5g mobile networks. *IEEE Wireless Communications*, 26(1):72–77.

[111] Liu Q, Han T, Ansari N, Wu G (2018) On designing energy-efficient heterogeneous cloud radio access networks. *IEEE Transactions on Green Communications and Networking*, 2(3): 721–734.

[112] Liu Y, Feng Q, Wu Q, Zhang Y, Jin M, Qiu T (2019) Energy-efficient hybrid precoding with low complexity for mmwave massive mimo systems. *IEEE Access*, 7:95021–95032.

[113] Luo Y, Shi Z, Bu F, Xiong J (2019) Joint optimization of area spectral efficiency and energy efficiency for two-tier heterogeneous ultra-dense networks. *IEEE Access*, 7:12073–12086.

[114] Mahapatra R, Nijsure Y, Kaddoum G, Hassan NU, Yuen C (2015) Energy efficiency tradeoff mechanism towards wireless green communication : A survey. *IEEE Communications Surveys & Tutorials*, 18(1):686–705.

[115] Mao Y, You C, Zhang J, Huang K, Letaief KB (2017) A survey on mobile edge computing : The communication perspective. *IEEE communications surveys & tutorials*, 19(4):2322–2358.

[116] Marabissi D, Bartoli G, Fantacci R, Micciullo L (2018) Energy efficient cooperative multicast beamforming in ultra dense networks. *IET Communications*, 12(5):573–578.

[117] Martínez F, Frías MP, Pérez MD, Rivera AJ (2019) A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review*, 52(3):2019–2037.

[118] Miozzo M, Giupponi L, Rossi M, Dini P (2017) Switch-on/off policies for energy harvesting small cells through distributed q-learning. *2017 IEEE wireless communications and networking conference workshops (WCNCW)*, IEEE, pages 1–6.

[119] Miozzo M, Piovesan N, Dini P (2019) Coordinated load control of renewable powered small base stations through layered learning. *IEEE Transactions on Green Communications and Networking*, 4(1):16–30.

[120] Misra R, Gudipati A, Katti S (2016) Quickc : practical sub-millisecond transport for small cells. *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 109–121.

[121] Morosi S, Piunti P, Del Re E (2013) Improving cellular network energy efficiency by joint management of sleep mode and transmission power. *2013 24th Tyrrhenian International Workshop on Digital Communications-Green ICT (TIWDC)*, IEEE, pages 1–6.

[122] Nasir YS, Guo D (2019) Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE Journal on Selected Areas in Communications*, 37(10):2239–2250.

[123] Ng DWK, Lo ES, Schober R (2012) Energy-efficient resource allocation in multi-cell ofdma systems with limited backhaul capacity. *IEEE Transactions on Wireless Communications*, 11(10): 3618–3631.

[124] Nguyen D, Tran LN, Pirinen P, Latva-aho M (2013) Precoding for full duplex multiuser mimo systems : Spectral and energy efficiency maximization. *IEEE Transactions on signal processing*, 61(16):4038–4050.

[125] Nguyen LD, Tuan HD, Duong TQ, Dobre OA, Poor HV (2018) Downlink beamforming for energy-efficient heterogeneous networks with massive mimo and small cells. *IEEE Transactions on Wireless Communications*, 17(5):3386–3400.

[126] Niu Z, Guo X, Zhou S, Kumar PR (2015) Characterizing energy–delay tradeoff in hyper-cellular networks with base station sleeping control. *IEEE Journal on Selected Areas in Communications*, 33(4):641–650.

[127] Niu Z, Wu Y, Gong J, Yang Z (2010) Cell zooming for cost-efficient green cellular networks. *IEEE communications magazine*, 48(11):74–79.

[128] Oh E, Krishnamachari B, Liu X, Niu Z (2011) Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Communications Magazine*, 49(6):56–61.

[129] Oh E, Son K, Krishnamachari B (2013) Dynamic base station switching-on/off strategies for green cellular networks. *IEEE transactions on wireless communications*, 12(5):2126–2136.

[130] Onireti O, Mohamed A, Pervaiz H, Imran M (2017) Analytical approach to base station sleep mode power consumption and sleep depth. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, pages 1–7.

[131] Orumwense EF, Afullo TJ, Srivastava VM (2016) Energy efficiency metrics in cognitive radio networks : A hollistic overview. *International Journal of Communication Networks and Information Security*, 8(2):75.

[132] Paleyes A, Urma RG, Lawrence ND (2022) Challenges in deploying machine learning : a survey of case studies. *ACM Computing Surveys*, 55(6):1–29.

[133] Park H, Lim Y (2018a) Energy-effective power control algorithm with mobility prediction for 5g heterogeneous cloud radio access network. *Sensors*, 18(9):2904.

[134] Park J, Samarakoon S, Bennis M, Debbah M (2018b) Wireless network intelligence at the edge. *arXiv preprint arXiv :1812.02858*.

[135] Parker MC, Nagraj S, Walker SD (2009) Absolute energy efficiency metric for carbon footprint resource management and network optimisation. *2009 International Conference on Clean Electrical Power*, IEEE, pages 111–116.

[136] Peesapati SKG, Olsson M, Masoudi M, Andersson S, Cavdar C (2021) Q-learning based radio resource adaptation for improved energy performance of 5g base stations. *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, IEEE, pages 979–984.

[137] Perabathini B, Baştuğ E, Kountouris M, Debbah M, Conte A (2015) Caching at the edge : A green perspective for 5g networks. *2015 IEEE international conference on communication workshop (ICCW)*, IEEE, pages 2830–2835.

[138] Perera TDP, Jayakody DNK, Sharma SK, Chatzinotas S, Li J (2017) Simultaneous wireless information and power transfer (swipt) : Recent advances and future challenges. *IEEE Communications Surveys & Tutorials*, 20(1):264–302.

[139] Piccinini G (2004) The first computational theory of mind and brain : a close look at mcculloch and pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, 141:175–215.

[140] Pompili D, Hajisami A, Tran TX (2016) Elastic resource utilization framework for high capacity and energy efficiency in cloud ran. *IEEE Communications Magazine*, 54(1):26–32.

[141] Popovski P, Simeone O, Boccardi F, Gündüz D, Sahin O (2020) Semantic-effectiveness filtering and control for post-5g wireless connectivity. *Journal of the Indian Institute of Science*, 100:435–443.

[142] Qiu Y, Zhang H, Long K, Huang Y, Song X, Leung VC (2018) Energy-efficient power allocation with interference mitigation in mmwave-based fog radio access networks. *IEEE Wireless Communications*, 25(4):25–31.

[143] Ramamoorthi Y, Kumar A (2017) Resource allocation for energy efficient next generation cellular networks. *CSI transactions on ICT*, 5:179–187.

[144] Ren P, Tao M (2014) A decentralized sleep mechanism in heterogeneous cellular networks with qos constraints. *IEEE Wireless Communications Letters*, 3(5):509–512.

[145] Rengarajan B, Rizzo G, Marsan MA (2015) Energy-optimal base station density in cellular access networks with sleep modes. *Computer Networks*, 78:152–163.

[146] Ribeiro LN, Schwarz S, Rupp M, de Almeida AL (2018) Energy efficiency of mmwave massive mimo precoding with low-resolution dacs. *IEEE Journal of Selected Topics in Signal Processing*, 12(2):298–312.

[147] Rodoplu V, Meng TH (2007) Bits-per-joule capacity of energy-limited wireless networks. *IEEE Transactions on Wireless Communications*, 6(3):857–865.

[148] Series M (2015) Imt vision–framework and overall objectives of the future development of imt for 2020 and beyond. *Recommendation ITU*, 2083(0).

[149] Shafin R, Liu L, Chandrasekhar V, Chen H, Reed J, Zhang JC (2020) Artificial intelligence-enabled cellular networks : A critical path to beyond-5g and 6g. *IEEE Wireless Communications*, 27(2):212–217.

[150] Sheng J, You Y, Ma D, Zhu C, Xu F (2018) Continuous cell zooming algorithm towards energy efficient in random heterogeneous cellular networks. *Communications and Networking : 12th International Conference, ChinaCom 2017, Xi'an, China, October 10–12, 2017, Proceedings, Part I 12*, Springer, pages 502–511.

[151] Sheng M, Wang L, Wang X, Zhang Y, Xu C, Li J (2016) Energy efficient beamforming in miso heterogeneous cellular networks with wireless information and power transfer. *IEEE Journal on Selected Areas in Communications*, 34(4):954–968.

[152] SMARTer2030 (2015) "ict solutions for 21st century challenges," available at http ://smarter2030.gesi.org/downloads/full report.pdf. *Tech. Rep., 2015. [Online]*.

[153] Soh YS, Quek TQ, Kountouris M, Shin H (2013) Energy efficient heterogeneous cellular networks. *IEEE Journal on selected areas in communications*, 31(5):840–850.

[154] Son K, Oh E, Krishnamachari B (2015) Energy-efficient design of heterogeneous cellular networks from deployment to operation. *Computer Networks*, 78:95–106.

[155] Suarez L, Nuaymi L, Bonnin JM (2012) An overview and classification of research approaches in green wireless networks. *Eurasip journal on wireless communications and networking*, 2012:1–18.

[156] Sun H, Zhou F, Hu RQ (2019) Joint offloading and computation energy efficiency maximization in a mobile edge computing system. *IEEE Transactions on Vehicular Technology*, 68(3):3052–3056.

[157] Tabassum H, Shakir MZ, Alouini MS (2012) Area green efficiency (age) of two tier heterogeneous cellular networks. *2012 IEEE Globecom Workshops*, IEEE, pages 529–534.

[158] Tan H, Feng Z (2018a) Power optimization in self-organizing mec based heterogeneous small cell networks. *IEEE Access*, 6:59109–59117.

[159] Tan W, Xie D, Xia J, Tan W, Fan L, Jin S (2018b) Spectral and energy efficiency of massive mimo for hybrid architectures based on phase shifters. *IEEE Access*, 6:11751–11759.

[160] Tang J, Shojaeifard A, So DK, Wong KK, Zhao N (2018) Energy efficiency optimization for comp-swipt heterogeneous networks. *IEEE Transactions on Communications*, 66(12):6368–6383.

[161] Tang J, So DK, Alsusa E, Hamdi KA, Shojaeifard A, Wong KK (2017) Energy-efficient heterogeneous cellular networks with spectrum underlay and overlay access. *IEEE Transactions on Vehicular Technology*, 67(3):2439–2453.

[162] Tao R, Liu W, Chu X, Zhang J (2019) An energy saving small cell sleeping mechanism with cell range expansion in heterogeneous networks. *IEEE Transactions on Wireless Communications*, 18(5):2451–2463.

[163] Tariq F, Dooley L, Allen B, Poulton A, Liu E (2017) A cell range expansion framework for closed access femtocell networks. *Wireless Personal Communications*, 93:601–614.

[164] Taufique A, Jaber M, Imran A, Dawy Z, Yacoub E (2017) Planning wireless cellular networks of future : Outlook, challenges and opportunities. *IEEE Access*, 5:4821–4845.

[165] Tombaz S, Frenger P, Athley F, Semaan E, Tidestav C, Furuskar A (2015) Energy performance of 5g-nx wireless access utilizing massive beamforming and an ultra-lean system design. *2015 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pages 1–7.

[166] Tung LP, Wang LC, Chen KS (2017) An interference-aware small cell on/off mechanism in hyper dense small cell networks. *2017 International conference on computing, networking and communications (ICNC)*, IEEE, pages 767–771.

[167] Venturino L, Zappone A, Risi C, Buzzi S (2014) Energy-efficient scheduling and power allocation in downlink ofdma networks with base station coordination. *IEEE transactions on wireless communications*, 14(1):1–14.

[168] Vu QD, Tran LN, Farrell R, Hong EK (2015) Energy-efficient zero-forcing precoding design for small-cell networks. *IEEE Transactions on Communications*, 64(2):790–804.

[169] Wang CX, Di Renzo M, Stanczak S, Wang S, Larsson EG (2020a) Artificial intelligence enabled wireless networking for 5g and beyond : Recent advances and future challenges. *IEEE Wireless Communications*, 27(1):16–23.

[170] Wang Q, Zhao F, Chen T (2018a) A base station dtx scheme for ofdma cellular networks powered by the smart grid. *IEEE Access*, 6:63442–63451.

[171] Wang X, Yang T, Cui Y, Jin Y, Wang H (2020b) Bsenet : A data-driven spatio-temporal representation learning for base station embedding. *IEEE Access*, 8:51674–51683.

[172] Wang X, Zhou Z, Xiao F, Xing K, Yang Z, Liu Y, Peng C (2018b) Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing*, 18(9):2190–2202.

[173] Weng X, Cao D, Niu Z (2011) Energy-efficient cellular network planning under insufficient cell zooming. *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, IEEE, pages 1–5.

[174] Wong WT, Yu YJ, Pang AC (2012) Decentralized energy-efficient base station operation for green cellular networks. *2012 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pages 5194–5200.

[175] Wu CH, Ho JM, Lee DT (2004) Travel-time prediction with support vector regression. *IEEE transactions on intelligent transportation systems*, 5(4):276–281.

[176] Wu J, Wong EW, Guo J, Zukerman M (2017) Performance analysis of green cellular networks with selective base-station sleeping. *Performance Evaluation*, 111:17–36.

[177] Wu J, Zhang Y, Zukerman M, Yung EKN (2015a) Energy-efficient base-stations sleep-mode techniques in green cellular networks : A survey. *IEEE communications surveys & tutorials*, 17(2):803–826.

[178] Wu J, Zhang Z, Hong Y, Wen Y (2015b) Cloud radio access network (c-ran) : a primer. *IEEE network*, 29(1):35–41.

[179] Wu YC, Chen K, Naglich EJ, Peroulis D (2013) A wideband 0.7–2.2 ghz tunable power amplifier with over 64% efficiency based on high-q second harmonic loading. *2013 IEEE MTT-S International Microwave Symposium Digest (MTT)*, IEEE, pages 1–4.

[180] Xiao L, Zhang H, Xiao Y, Wan X, Liu S, Wang LC, Poor HV (2019) Reinforcement learning-based downlink interference control for ultra-dense small cells. *IEEE Transactions on Wireless Communications*, 19(1):423–434.

[181] Xu J, Duan L, Zhang R (2015) Cost-aware green cellular networks with energy and communication cooperation. *IEEE Communications Magazine*, 53(5):257–263.

[182] Xu J, Ota K, Dong M (2018) Saving energy on the edge : In-memory caching for multi-tier heterogeneous networks. *IEEE Communications Magazine*, 56(5):102–107.

[183] Xu K, Van Huynh N, Li GY (2023) Distributed-training-and-execution multi-agent reinforcement learning for power control in hetnet. *IEEE Transactions on Communications*.

[184] Xu X, Yuan C, Chen W, Tao X, Sun Y (2017) Adaptive cell zooming and sleeping for green heterogeneous ultradense networks. *IEEE Transactions on Vehicular Technology*, 67(2):1612–1621.

[185] Yang C, Li J, Ni Q, Anpalagan A, Guizani M (2016) Interference-aware energy efficiency maximization in 5g ultra-dense networks. *IEEE Transactions on Communications*, 65(2):728–739.

[186] Yang H, Li X, Qiang W, Zhao Y, Zhang W, Tang C (2021) A network traffic forecasting method based on sa optimized arima–bp neural network. *Computer Networks*, 193:108102.

[187] Yang T, Héliot F, Foh CH (2018) Energy-efficient boundary-enabled scheduling in the downlink of multi-carrier multi-access heterogeneous network. *IEEE Transactions on Green Communications and Networking*, 3(1):79–92.

[188] Yang Y, Gao F, Xing C, An J, Alkhateeb A (2020) Deep multimodal learning : Merging sensory data for massive mimo channel prediction. *IEEE Journal on Selected Areas in Communications*, 39(7):1885–1898.

[189] Ye J, Gharavi H (2020) Deep reinforcement learning-assisted energy harvesting wireless networks. *IEEE transactions on green communications and networking*, 5(2):990–1002.

[190] Ye Q, Rong B, Chen Y, Al-Shalash M, Caramanis C, Andrews JG (2013) User association for load balancing in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, 12(6):2706–2716.

[191] Yi B, Wang X, Li K, Huang M et al. (2018) A comprehensive survey of network function virtualization. *Computer Networks*, 133:212–262.

[192] Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks : Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

[193] Zappone A, Jorswieck EA, Buzzi S (2014) Energy efficiency and interference neutralization in two-hop mimo interference channels. *IEEE Transactions on Signal Processing*, 62(24):6481–6495.

[194] Zappone A, Sanguinetti L, Bacci G, Jorswieck E, Debbah M (2015) Energy-efficient power control : A look at 5g wireless technologies. *IEEE Transactions on Signal Processing*, 64(7): 1668–1683.

[195] Zeng M, Hao W, Dobre OA, Poor HV (2019) Energy-efficient power allocation in uplink mmwave massive mimo with noma. *IEEE Transactions on Vehicular Technology*, 68(3):3000–3004.

[196] Zeng M, Yadav A, Dobre OA, Poor HV (2018) Energy-efficient power allocation for mimo-noma with multiple users in a cluster. *IEEE Access*, 6:5170–5181.

[197] Zhai D, Zhang R, Du J, Ding Z, Yu FR (2018) Simultaneous wireless information and power transfer at 5g new frequencies : Channel measurement and network design. *IEEE Journal on Selected Areas in Communications*, 37(1):171–186.

[198] Zhang C, Patras P, Haddadi H (2019a) Deep learning in mobile and wireless networking : A survey. *IEEE Communications surveys & tutorials*, 21(3):2224–2287.

[199] Zhang C, Zhang H, Yuan D, Zhang M (2018) Citywide cellular traffic prediction based on densely connected convolutional neural networks. *IEEE Communications Letters*, 22(8):1656–1659.

[200] Zhang H, Liu H, Cheng J, Leung VC (2017a) Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks. *IEEE transactions on communications*, 66(4):1705–1716.

[201] Zhang T, Zhao J, An L, Liu D (2016) Energy efficiency of base station deployment in ultra dense hetnets : A stochastic geometry analysis. *IEEE Wireless Communications Letters*, 5(2): 184–187.

[202] Zhang Y, Cui Z, Cui Q, Yu X, Liu Y, Xie W, Zhao Y (2017b) Energy efficiency analysis of heterogeneous cellular networks with extra cell range expansion. *IEEE Access*, 5:11003–11014.

[203] Zhang Z, Liu F, Zeng Z (2017c) The cell zooming algorithm for energy efficiency optimization in heterogeneous cellular network. *2017 9th International conference on wireless communications and signal processing (WCSP)*, IEEE, pages 1–5.

[204] Zhang Z, Xiao Y, Ma Z, Xiao M, Ding Z, Lei X, Karagiannidis GK, Fan P (2019b) 6g wireless networks : Vision, requirements, architecture, and key technologies. *IEEE vehicular technology magazine*, 14(3):28–41.

[205] Zhao Z, Chen W, Wu X, Chen PC, Liu J (2017) Lstm network : a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75.

[206] Zheng J, Gao L, Zhang H, Zhu D, Wang H, Gao Q, Leung VC (2018) Joint energy management and interference coordination with max-min fairness in ultra-dense hetnets. *IEEE Access*, 6:32588–32600.

[207] Zheng Y, Sun S, Rong B, Kadoch M, Yamao Y (2015) Traffic aware power allocation and frequency reuse for green lte-a heterogeneous networks. *2015 IEEE International Conference on Communications (ICC)*, IEEE, pages 3167–3172.

[208] Zheng Z, Gharavi H (2019) Spectral and energy efficiencies of millimeter wave mimo with configurable hybrid precoding. *IEEE transactions on vehicular technology*, 68(6):5732–5746.

[209] Zhou L, Sheng Z, Wei L, Hu X, Zhao H, Wei J, Leung VC (2016) Green cell planning and deployment for small cell networks in smart cities. *Ad Hoc Networks*, 43:30–42.

[210] Zhou T, Jiang N, Liu Z, Li C (2017) Joint cell activation and selection for green communications in ultra-dense heterogeneous networks. *IEEE Access*, 6:1894–1904.