Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Hybrid Clustering and Deep Learning-based Spatio Temporal Analysis of Spectrum Utilization

By

Frehiwot Bantigegn

Advisor

Dr. -Ing. Dereje Hailemariam

A Thesis Submitted to the School of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecom Network Engineering

October 2021

Addis Ababa, Ethiopia

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Hybrid Clustering and Deep Learning-based Spatio Temporal Analysis of Spectrum Utilization

By

Frehiwot Bantigegn

Approval by Board of Examiners

_____	_____
Chairman, School Graduate Committee	Signature
Dr. -Ing. Dereje Hailemariam	_____
Advisor	Signature
_____	_____
Internal Examiner	Signature
_____	_____
External Examiner	Signature

Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in this or any other university, and all sources of materials used for the thesis have been fully acknowledged.

Frehiwot Bantigegn _____

Name Signature

Place: Addis Ababa, Ethiopia

Date of Submission:_____

As a university advisor, I approved the submission of this thesis for examination.

Dr.-Ing. Dereje Hailemariam _____

Name Signature

Abstract

Radio spectrum is a finite resource, while the demand for wireless systems is increasing at an exponential rate. To meet this demand, new generations of cellular networks were introduced.Spectrum utilization of cellular bands is analyzed widely using spectrum measurements. Knowledge of spectrum utilization will help operators like Ethio telecom to understand and plan band usage.

In this thesis, using the K-means algorithm and Deep learning algorithms, namely Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), downlink Global System for Mobile Communication (GSM) 900 spectrum utilization is analyzed and modeled to know the spectrum utilization of Ethio telecom. The data is collected from Addis Ababa 639 GSM base stations. Spectrum utilization is modeled using CNN and LSTM algorithms for clustered and non-clustered data. Because of the differences in base station behavior, clustering base stations is done and model the spectrum utilization of the base stations in each cluster.

Our results show that the GSM 900 downlink spectrum is not utilized optimally. The highest observed average spectrum utilization was 71%, with the lowest observed average spectrum utilization being 1.4%. The model developed for the cluster data using the CNN algorithm can model spectrum utilization with an RMSE value of 0.58 and this model can predict the next twenty-four-hour base station spectrum utilization with an RMSE value of 1.04.

Keywords – Spectrum Utilization, GSM900, Downlink, K-means, LSTM, CNN.

Table of Contents

List of Figures

List of Tables

Dedication

To My Beloved Parents: Bantigegn Melese and Dejitnu Tamene

To My Lovely Husband: Binyam Derbie

To My Little Daughter: Ermakel Binyam

Acknowledgment

First and foremost, I want to express my gratitude to the Almighty God and to his mother, Holy Virgin Mariam, for being my strength and guidance throughout my life, as well as for giving me so many blessings.

I would like to express my deep gratitude to my adviser, Dr. –Ing. Dereje Hailemariam, for his dedication and unwavering support, guidance, encouragement, and suggestions. During my research, his patience with me was amazing. Without his time, constructive feedback, and follow-up, this research would not have been possible. I am also very grateful to thank Mrs. Bethlehem Seifu for her support and Ethio telecom staff Mr. Gizachewu Addis and Mr. Dawit Kebede, who supported me in gathering data and providing necessary information for the effective completion of this thesis.

I would also like to thank my examiners Dr. Beneyam Berehanu and Dr. Eng. Yihenew Wondie for their valuable comments and suggestions.

Abbreviation

2G	Second Generation
3G	Third Generation
4G	Fourth Generation
5G	Fifth Generation
AI	Artificial Intelligence
AuC	Authentication Center
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base transceiver station
CNN	Convolution Neural Network
CS	Circuit Switching
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCS	Digital Cellular Radio System
DSA	Dynamic Spectrum Allocation
ECA	Ethiopian Communication Authority
EIR	Equipment Identity Register
FDMA	Frequency-division multiple access

GoS	Grade of Service
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
HLR	Home Location Register
ITU	International Telecommunication Union
LSTM	Long short-term memory
LSVM	Lagrangian Support Vector Machine
LTE	Long Term Evolution
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MS	Mobile Station
MSC	Switching Services Center
MSE	Mean Square Error
NSS	Network and switching subsystem
OSS	Operation and support subsystem
PRS	Performance Report System
RAT	Radio Access Technology
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
TCH	Traffic channel
TCH/F	Full Rate Traffic Channel
TCH/H	Half Rate Traffic Channel
TDMA	Time Division Multiple Access

TRX Transmit and Receive

UMTS Universal Mobile Telecommunications System

VLR Visitor Location Register

Chapter 1

Introduction

Radio spectrum is a finite resource while the demand for wireless systems is increasing at an exponential rate. Operators and regulators are examining ways to increase the available spectrum and efficient utilization of the existing spectrum [1].

The radio spectrum is defined by three dimensions: space, frequency, and time. Spectrum utilization refers to the amount of information that is being carried by a spectrum unit. The measures of spectrum utilization are divided into two categories. Namely, absolute measures and relative measures. Absolute measures can be used for assessing the overall level of spectrum utilization. It is defined as being the product of the frequency bandwidth, geographic space, and time. Relative measures, on the other hand, apply to specific services with known network parameters [2], [3]. Ethio telecom system measures channel utilization from traffic volume and assigned channel numbers.

Measurement-based radio spectrum utilization assessment is one important method to provide operators, regulators, and researchers with information regarding current, and future spectral utilization, both for incumbent and new users [4].

For network operators such as Ethio telecom, knowing the spectrum utilization will help to understand and plan band usage and for the resource allocation process. Ethio telecom implements different methods to increase spectrum utilization. Refarming at 900MHz and 1800MHz, implementing half-rate channel configuration, and load balancing using parameter optimizations are some.

Spectrum utilization can be characterized in frequency, time, and/or space aspects [3]. In this thesis work, spatial and temporal spectrum utilization analysis and base station level prediction is done. Data is obtained from Ethio telecom performance report systems for 636 Global System for Mobile communication (GSM) base stations for 100 days in Addis Ababa. Spectrum utilization is modeled as time series data. In this thesis, deep neural models are used to increase this model's accuracy. Ethio telecom will be able to use these results as information for frequency planning and management. It is also used to allocate the right resources at the right place.

1.1 Statement of the Problem

Spectrum utilization studies have been performed to know the exact behavior of the allocated spectrum in specific environments. Spectrum utilization analysis can be used to assess the current status of spectrum use and the availability of the spectrum for others to use. An appropriate study of current spectrum usage help to know unoccupied spectral resources in terms of frequency, time, and space. Therefore, the measurement of real network spectrum utilization creates an important step towards a realistic understanding of effective spectrum utilization [5].

As shown in Table 1.1, In Ethio telecom static spectrum allocation is done for cellular networks. Around 70% (i.e., 17.5MHz) of the 900 MHz spectrum band is configured for GSM and Universal Mobile Telecommunications System (UMTS) services. From the assigned 17.5MHz for GSM 900 service 12.5 MHz is configured. This is the reason for this research to target the GSM 900 MHZ spectrum utilization. Moreover, Ethiopian Communication Authority (ECA) has a new band plan for Ethio telecom and new entrants.

Table 1.1: Spectrum Usage in Ethio telecom

Band	Available BW	Current Usage	Use
900MHz	25MHz X 2	17.5MHz	UMTS + GSM
1800MHz	75MHz X 2	37.5MHz	LTE + GSM
2100MHz	60MHz X 2	20MHz	UMTS
2600 MHz	70Mhz X 2	20MHz	LTE

1.2 Objective

1.2.1 General Objective

The overall goal of this thesis is to analyze and model spectrum utilization using k-means clustering and deep learning methods on data obtained from Ethio telecom GSM networks.

1.2.2 Specific Objective

The following specific objectives have been identified in order to meet the study's overall goal.

- Review literature
- Collect the data from the performance report system (PRS)
- Applying the required data preprocessing techniques on the data set
- Study clustering algorithms and using the K-means algorithm clustering the base stations according to their behavior to analyze the spectrum usage behavior.
- Model spectrum utilization of Addis Ababa GSM network using deep learning algorithms.
- Clustering before applying the forecasting techniques and examining the effect of clustering on the accuracy of the predicted results.
- Evaluate the performance of the model using the performance evaluation metrics and select the best model for spectrum utilization analysis and base station level prediction.
- Result analysis and conclusion

1.3 Literature Review

Spectrum occupancy information is necessary for a cognitive radio network (CRN). It also helps in modeling and predicting the spectrum availability for efficient dynamic spectrum access (DSA). DSA systems typically consist of licensed primary users and opportunistic secondary users. Primary users are the present owners of the spectrum, while the secondary users opportunistically access the spectrum [6].

Different works of literature have used different approaches to study spectrum utilization of different bands. In most literature spectrum utilization measurement and analysis are studied based on the concepts of dynamic spectrum access in the context of cognitive radio. Some literature related to spectrum utilization measurement and analysis is reviewed in this section.

Motivated by the lack of knowledge regarding spectrum occupancy in South Africa, the authors in [7] measured the spectrum occupancy in the ultra-high frequency (UHF), GSM 900 MHz, and 1800 MHz bands. Measurements were taken on a daily basis, spaced at two hourly intervals for six weeks, and with a sampling resolution of 2Mhz for UHF and 100 kHz for the GSM bands. The results indicate a maximum occupancy of 20% for UHF bands. For the GSM bands, during peak hours the maximum utilization for the GSM 900 MHZ and 1800 MHz bands are 92% and 40%, respectively.

In [8] GSM channel utilization modeling and prediction are done. Spectrum measurements were performed in Bogota City, Colombia, in the GSM band of 850MHz.For 7 days, 60 channels were measured in this band. From the measured channels, three channels were chosen. The chosen channels had high, medium, and low occupancy levels. To model and predict the channel utilization Seasonal Autoregressive Integrated Moving Average (SRIMA) has been used. The result of primary users (PUs) occupancy models can be used as an empty channel. Besides that, based on prediction information, SUs can select the channels with a higher probability of availability in multi-channel wide band sensing scenarios .

The study in [6] analyzes the practical prowess of time-series modeling methodologies, such as Autoregressive (AR) and Auto Regressive Integrative Moving Average (ARIMA)

models, and machine learning techniques, such as Lagrangian Support Vector Machine (LSVM) and simplified models of RNNs, i.e., Elman network (EN), for predicting spectrum for a spectrum measurement in Jaipur, Rajasthan, India. The measurements were performed for one week. The frequency range was 150-750 MHz in the TV band and 850-1 300 MHz in the cellular band. The performance evaluation is done using MSE and the study discovers that due to cellular data traffic is non-stationary with several irregularities, the RNN technique outperforms the other model in terms of prediction accuracy. While in the TV band, the traffic pattern is stationary, and the time-series models can work efficiently.

Paper [9] addresses the problem of inefficient spectrum utilization in GSM using dynamic spectrum sharing (DSS) between mobile network operators. The proposed spectrum sharing scheme considers spectrum utilization and calls blocking probability. It is evaluated under the different traffic conditions for base stations. The result for the proposed scheme shows improvement in spectrum utilization with reduced call blocking probability.

To summarize, due to their capacity to capture the spectrum utilization data set's non-stationary behavior deep learning models are recommended for spectrum usage prediction over time series prediction methods [6] like the one used in this thesis. To the best of my knowledge so far, no research has been done for analyzing Ethio telecom spectrum utilization using the clustering and deep learning algorithms.

1.4 Methodology

The methodology begins with reading various papers, journals, articles, and books on the spectrum and measurement techniques. The data set used in this thesis was gathered from the Ethio telecom GSM performance report system on an hourly basis for 100 days. For the collected data sets, the required data preprocessing and data analysis techniques are applied. Using the preprocessed data, clustering is done using the K-means algorithm. The clustered data is used for model building as clustering can be used as one preprocessing technique. Model building is also done for clustered and non-clustered data. Then, model performance evaluation methods are applied to select the better model.

Microsoft Excel, Microsoft Visio, Python, Keras, and TensorFlow libraries are used to simulate and analyze the study's results.The overall methodology for this thesis is as follows:

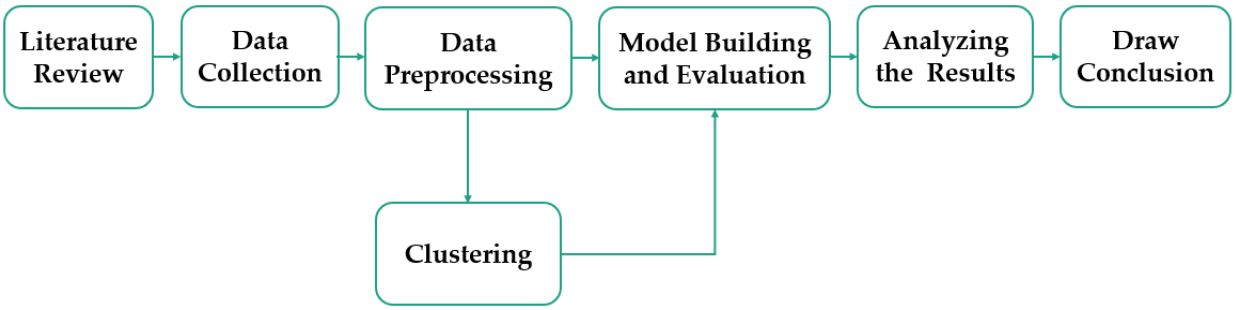

Figure 1.1: Overall Methodology

1.5 Scope and Limitation

1.5.1 Scope of the Thesis

In this thesis, a model based on deep learning is developed. The modeling considers solely voice traffic on the GSM 900 network in Addis Ababa, as data traffic on the GSM network is insignificant.

1.5.2 Limitation of the Thesis

The limitation of this thesis is that only the GSM 900 network of the Ethio telecom network in Addis Ababa is examined. But a similar approach can be followed for GSM 1800 as well as for different coverage areas of the operator.

1.6 Contribution

This research could be the starting point for researching managing spectrum utilization in Ethio telecom. Moreover, the contributions of the thesis are as follows:

- This research outputs provide significant insight to Ethio telecom's growing spectrum demand as it will be ending of monopoly helps to plan the effective spectrum usage strategy by taking this spectrum utilization result as an input.
- It can be used to improve parameters of load balancing algorithm
- It can be used to assess the cost of adding more TRX when the traffic can be handled with fewer TRX.
- It can be an input for better spectrum utilization through dynamic spectrum allocation and also used to plan an energy-efficient network.

1.7 Thesis Organization

This thesis is organized into five chapters. The first Chapter is a thesis introduction. The second Chapter defines the spectrum, which frequency bands are assigned to cellular networks, and assignments in Ethio telecom, and gives an overview of GSM technologies. The third Chapter discusses the fundamentals of clustering, deep learning algorithms used in this thesis, hyperparameter tuning methods, and model evaluation metrics. In Chapter four, Ethio telecom spectrum usage analysis, the data set used for the paper and the techniques used for analyzing the utilization, model building, model evaluation, and results obtained are discussed. The last Chapter concludes the results obtained from this thesis and recommends future works that are not included in this thesis.

Chapter 2

Spectrum Allocation in Cellular Networks

2.1 Introduction

Spectrum is a national resource. Each country has complete authority over how it is used. The International Telecommunication Union (ITU) brings governments together to assign specific bands to specific services on a global or regional scale [9]. There are two types of spectrum control: licensed and unlicensed. Most radio spectrum is licensed and covers a range of technologies that operate with suitable power to allow the services to cover a relatively wide area. National regulators manage access to this spectrum through a licensing framework. It allows an organization the exclusive rights to use a certain frequency band in certain areas and at certain times. Unlicensed frequency bands have more limited applications and are designated for certain specific types of use. A license from the regulator is not required as long as the devices used to meet certain technical standards to minimize interference [1].

Radio spectrum relates to the invisible radio frequencies assigned to the mobile industry and other areas for communication. Depending on their frequency, they can pass through solid objects and travel long distances. This makes them useful for mobile communications, broadcasting and many other wireless applications.Radio spectrum is a finite resource while the demand for wireless systems is increasing at an exponential rate [10].

Spectrum resource describes the availability of spectrum in terms of space, time, and a number of channels that all users on a certain territory may access. In cellular systems, the spectrum resource may consist of all frequency channels in a certain band but may be limited in time [3].

Radio spectrum is divided into frequency bands. The band is subdivided into channels that are used for transmission. Not all radio frequencies are equal. In general, lower frequency bands provide wider coverage because they can penetrate objects effectively and thus travel further, including inside buildings. However, they tend to have relatively poor capacity.Higher frequency bands do not provide as good coverage as the signals are weakened or even stopped by obstacles such as buildings [10].

To meet the increasing demand for high capacity, high data rate, and low latency, in the past two decades, the cellular networks have evolved different generations, from first generation network to the second generation network to the third generation network to the fourth generation network, and fifth generation network. When a new generation network is introduced, the mobile traffic will gradually migrate from the older generation (i.e., 2G/3G) networks towards the new one. For the period of generation change, the spectrum of the old network will face low spectral efficiency when the number of users is lower than the designed network capacity [11].

2.2 Frequency Bands in Cellular Mobile Networks

Spectrum is one of the most demanded resources for mobile telecommunications operators. Spectrum policy is vital as spectrum is a scarce resource allocated by regulatory authorities. Different spectrum bands, such as 800 MHz, 900 MHz, 1800 MHz, 2100 MHz, and 2600 MHz, are used by mobile telecommunications services. The propagation characteristics of the different bands vary. This variation has an impact not only on costs for network roll out but also on the relative value of the spectrum price [12].

Spectrum band assigned to telecom companies is extremely important. They pay switching expenses when migrating from one frequency band to another. Telecom operators have built a grid of base stations based on the propagation characteristics of the frequency

bands they have been allocated. When analyzing the spectrum allocated to an operator the distinction between spectrum below 1GHz and spectrum above 1GHz is important. The coverage spectrum below 1GHz has a lower propagation path loss than a spectrum above 1 GHz [13].

Most operators around the world use the 900 MHz and 1800 MHz bands for providing 2G mobile services using GSM technology. In GSM 900 MHz the uplink frequency band is 890–915 MHz, and the downlink frequency band is 935–960MHz [14].

As 3G services have come to dominate mobile networks in recent years, many operators now find this 900 MHz spectrum underutilized and reframing 5MHz band width of 900MHz to UMTS. Refarming is the term used for the process governing the repurposing of frequency bands that have historically been allocated for 2G mobile services (using GSM technology) for new generation of mobile technologies, including both third generation (using UMTS technology) and fourth generation (using Long term Evolution (LTE) technology) [14].

Paper [15] has assessed the differences in costs between networks using different spectrum bands. The authors have analyzed the cost savings for UMTS900 compared to 2100MHz.Based on an economic assessment of costs, they conclude that there are significant benefits from the roll-out of a UMTS900 compared to a UMTS2100 MHz network.

The 1800MHz has the richest spectral resources, as its bandwidth is 75MHz.This band is widely used by operators for 2G GSM services, it provides a cost-effective solution for boosting mobile broadband capacity—allowing operators to refarm spectrum they already own for LTE, instead of (or in addition to) licensing new spectrum refarming 1800MHz for LTE is the most feasible and cost-efficient way for operators to provide high-speed data services and allows early market entry [14].

2.3 Frequency Band Allocation in Ethio telecom

Operators around the world, similarly Ethio telecom already use the 1800 MHz and 900MHz band for 2G GSM services. As 3G and 4G services have come to dominate

mobile networks in recent years, many operators including Ethio telecom refarm 900Mhz and 1800 MHz spectrum bands for UMTS and LTE services, respectively.

As shown in Figure 2.1 and 2.2 Ethio telecom refarmed 5MHz from 900Mhz and 20Mhz from 1800Mhz for UMTS and LTE services. Generally, Ethio telecom deployed GSM900, GSM1800, UMTS900, UMTS2100, LTE1800 and LTE2600 sites to support the data and voice traffic demand of Addis Ababa.

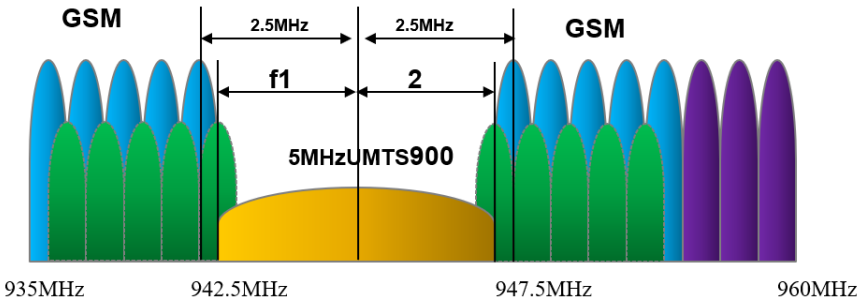

Figure 2.1: Ethio telecom UMTS 900 Refarming [16]

In Ethio telecom from 25 MHz ranging from 935-960MHz frequency range from 947.5 -960MHz(12.5Mhz) is configured for GSM service and frequency 942.5 -947.5 (5MHz) is refarmed to UMTS 900.In Ethio telecom 20Mhz band width from a frequency range,1851.5 to 1871.5 is refarmed to LTE1800. 2100 MHz, and 2600MHz bands are mainstream for UMTS and LTE, respectively.

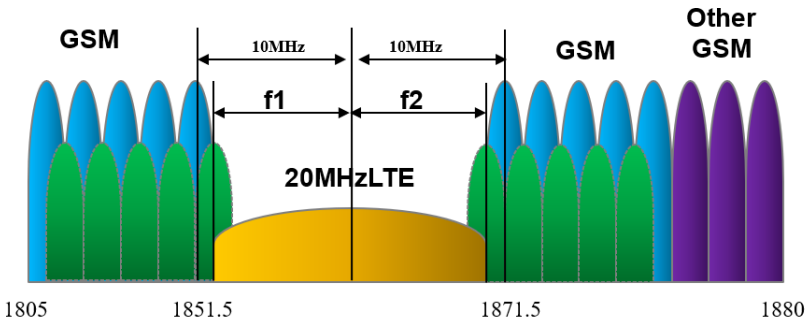

Figure 2.2: Ethio telecom LTE 1800 Refarming [16]

2.4 Introduction to GSM System

2.4.1 GSM System Architecture

The GSM network is structured hierarchically and grouped into four main areas as shown in Figure 2.3 [17].

A Mobile station (MS) is the component of a GSM cellular network that the user sees and operates.

The Base-station subsystem (BSS) is associated with communicating with the mobiles on the network. It is composed of the base stations (BTS's) and the base station controllers (BSC's). The BSS uses the Abis interface between the BTS and the BSC.

- **Base transceiver station (BTS)**: The BTS used in a GSM network consist of the radio transmitter-receivers, and their associated antennas that transmit and receive to directly communicate with the mobiles. Transmission execution, channel encryption, diversity, and frequency hopping are all performed in this section. The Um interface, with its associated protocols, is the interface between the BTS and MS.
- **Base station controller (BSC)**: It controls a group of BTSs. It manages the radio resources and controls items such as handover within the group of BTSs and allocates channels. It communicates with the BTSs over the Abis interface.

Network and switching subsystem (NSS): the main part of which is the Mobile Switching Center (MSC), which performs the switching of calls, as well as the management of mobile services such as authentication. The major elements within the core network and include mobile switching services center (MSC), home location register (HLR), visitor location register (VLR), equipment identity register (EIR), authentication center (AuC).

Operation and support subsystem (OSS) The Operation and Support Subsystem (OSS) is an element within the overall GSM network architecture that is connected to components of the NSS and the BSC. It is used to control and monitor the overall GSM network and it is also used to control the traffic load of the BSS [18].

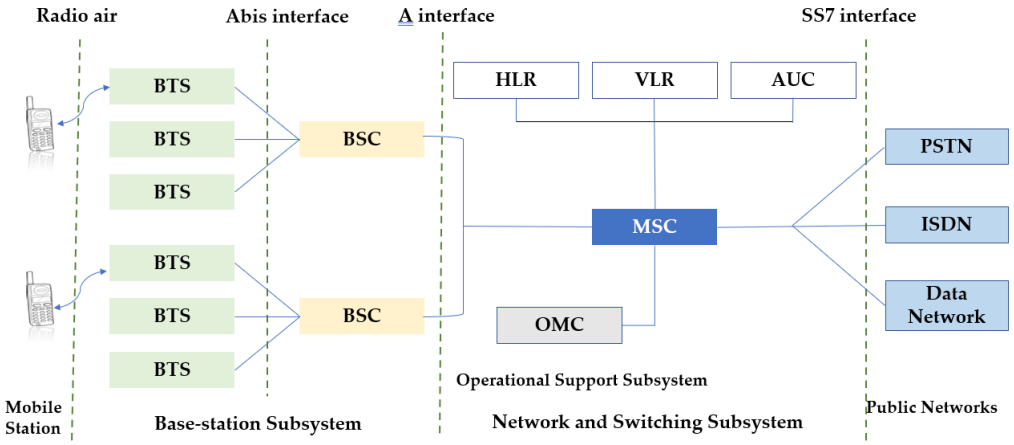

Figure 2.3: Network Architecture [17]

2.4.2 Multiple Access Scheme in GSM

The radio spectrum is a scarce resource. Therefore, GSM technology uses TDMA and FDMA mechanisms to allow multiple users to access the same radio spectrum at the same time.In GSM, the multiple access scheme is based on the multi-carrier, time division multiple access, and frequency division duplex, MC/TDMA/FDD principle. Two frequency bands are defined for GSM900: the band of 890-915 MHz is used for the uplink and the band of 935-960 MHz is used for the downlink. These bands are, in most countries, divided among two or three operators.

In addition to the 900 MHz band, the 1800 MHz band is also available for GSM service, ranging from 1710 MHz to 1785 MHz for the uplink and from 1805 MHz to 1880 MHz for the downlink. The carrier spacing is in both cases 200 kHz allowing 124 carriers in GSM900 and 374 carriers in GSM1800. The channel separation between uplink and downlink is 45MHz in the case of GSM 900 and is 95 MHz in the case of the DCS network. Each radio frequency is divided into TDMA frames of 4.615 ms, with each TDMA frame subdivided into 8 full timeslots. Each of these timeslots can carry a full rate traffic channel, two half-rate traffic channels, or one of the control channels. One timeslot on one frequency is called a slot [19].

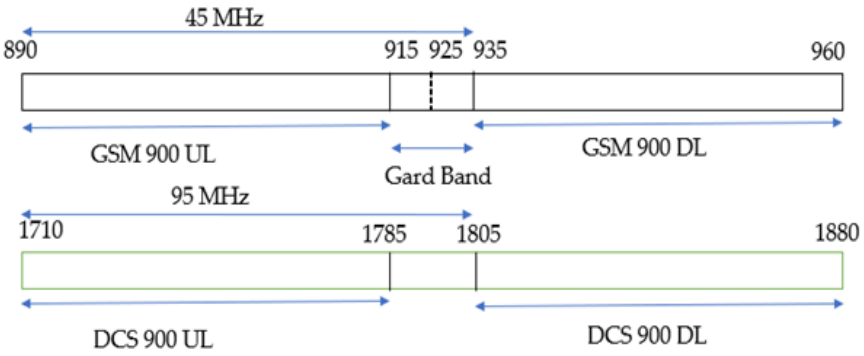

Figure 2.4: Channel Assignment

2.4.3 GSM Channels

Channels are described on two levels: the physical and the logical level. A physical channel corresponds to one timeslot on one carrier, while a logical channel reflects the specific type of information carried by the physical channel. This means that the different sorts of information will be sent on different logical channels. Typically, the logical channels are divided into two groups: traffic channels and control channels.

Traffic channels are the resources available to the user for either speech or data. A logical traffic channel in GSM is abbreviated by TCH. It can be used for either data or speech. The radio interface must support bi-directional transmission in order to support speech and data communication. For voice, 9.6 kbit/s and 4.8 kbit/s can be supported on full, and half rate traffic channels.

Control channels are the channels used for signaling and controlling. In other words, they control the traffic channels. In GSM there are common as well as dedicated control channels. The broadcast channels (BCH) are only used in the DL, they provide the MS with the information needed to establish the synchronization on both time and frequency, and broadcast control channel (BCCH), synchronization channel (SCH) and frequency correction channel (FCCH) are BCH channels. BCCH is the most important control channel within the GSM. A mobile station can receive, several cells. It has then to choose one of them, and some informations like the network to which each cell belongs. This information is broadcast regularly in each cell, to be listened to by all the mobile stations in idle mode. The common control channels (CCCH) are used to send information to a certain MS to initiate the setup stage before a channel is allocated to that MS, and paging

channel (PCH), access grant channel (AGCH) and random-access channel (RACH) are CCCH channels. The dedicated control channels (DCCH) are bidirectional and transmit the signaling information that is necessary during a connection, such as the assurance that BS and MS stay connected during the authentication process, the information update of the signal quality received at the MS, or handover procedures [20].

2.4.4 Traffic Engineering

Traffic engineering is fundamental to the design of circuit-switched networks. The traffic-performance relation is given by the Erlang loss formula which gives the probability of call blocking when a certain volume of traffic is offered to a given number of circuits [21].

Cellular radio systems depend on trunking to accommodate a large number of users in a limited radio spectrum. Trunking allows a large number of users to share the relatively small number of channels in a cell by providing access to each user, on demand, from a pool of available channels. In a trunked radio system, each user is allocated a channel on a per call basis, and upon termination of the call, the previously occupied channel is immediately returned to the pool of available channels [22]. Below, there is a brief definition of some parameters used to configure channels for trunked radio systems that can handle a specific capacity.

Grade of service (GOS) measure of the call blocking in voice traffic, where it measures of the ability of a user to access a trunked system during the busiest hours.
Traffic intensity offered by each user is equal to the call request rate multiplied by the holding time. Each user generates a traffic intensity of AU Erlangs given by

$$A_U = \lambda * H \tag{2.1}$$

where H is the average duration of a call and lambda is the average number of call requests per unit time. If a system has U users and an unspecified number of channels, the total offered traffic intensity A, is given as

$$A = U * Au \tag{2.2}$$

If the traffic is equally distributed among the channels, then the traffic intensity per channel, is shown in equation 2.3, where C is channels trunked system.

$$A_C = (U * A_U) / C \tag{2.3}$$

Advanced Mobile Phone Service(AMPS) cellular system is designed for a GOS of 2 percent blocking [22]. This implies that the channel allocations for cell sites are designed so that 2 out of 100 calls will be blocked due to channel occupancy during the busiest hour.

Chapter 3

Basics of Clustering and Deep Learning Algorithms

3.1 Time Series Analysis

The term time series refers to a sequence of data points. They are assembled over even intervals in time and ordered chronologically. Time series analysis includes methods for analyzing time series data to extract meaningful characteristics of the data and forecast future values. Auto correlation, trend, noise, and seasonal variation are characteristics of time series data. Time series data can either be univariate when the recorded sequence is for a single observation or multivariate when it is for multiple sets of observations [22]. Spectrum utilization is modeled as a time series model.

3.1.1 Components of Time Series Data

Some or all of the following components can be found in time series data [23].

The Trend: is the long-term pattern of a time series. Depending on whether the time series shows a growing or declining long-term pattern, a trend might be positive or negative.

The Cyclical component: shows an up and down movement around the trend.

A Seasonal component: is concerned with fluctuations that occur on a regular basis.

The Irregular component:is unpredictable. Every time series contains some unpredictability, making it a random variable.

3.2 Clustering Algorithms

Clustering is the process of grouping the data points into some groups. The similarity between the data points lying in the same cluster is higher than the similarity between them and the data points lying in the other group. Clustering is widely used in many applications, such as business intelligence, market research, image pattern recognition, data analysis, Web search, biology, and security [24].

Some clustering techniques [25]:

- **Partitioning Method:** For a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster. That is, it divides the data into K groups, which must contain at least one object. It performs one-level partitioning on data sets and adopts exclusive cluster separation, which means each object must belong to exactly one group. E.g., K-means
- **Hierarchical Method:** A hierarchical method creates a hierarchical decomposition of the given set of data objects. E.g., Agglomerative clustering, bottom-up, and Divisive hierarchical clustering, from top to bottom.
- **Density-based:** The dense regions among data samples to form clusters and low-density regions create boundaries between the clusters. DBSCAN algorithm popular density-based algorithm.

According to [26] five unsupervised clustering methods: k-means, hierarchical clustering, DBSCAN, spectral clustering and Birch are compared. Among these methods, K-means turns out to be the most suitable algorithm to cluster large data sets in terms of run-time and accuracy.

3.2.1 K-means Algorithm

The K-means algorithm is a partition clustering algorithm. The algorithm is based on unsupervised learning used with unlabeled multidimensional data. The goal of the algorithm is to group the unlabeled multidimensional data into K clusters. The K variable represents the number of groups for the partition. It works by iteratively assigning data points to one of the K groups based on the provided features. Each data point is assigned to one unique group. The algorithm is favored in many application areas such as computer vision, image processing, business analytics, etc. Its popularity is due to its simplicity and linear complexity [27].

The goal of K-means is creating clusters with highest possible similarity between the data inside each cluster and minimum similarity between the data in different clusters.The objective of K-means clustering is to minimize total intra-cluster variance, or the squared error function [28].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - C_j \right\|^2 \tag{3.1}$$

Where J is objective function, k is number of cluster, n is number of data points, and $\left\| x_i^{(j)} - C_j \right\|^2$ is the distance metric used in the algorithm.

Euclidean Distance is metrics used to measure the similarity between two time series and it is a well -known clustering distance measure.
Euclidean distance is calculated as:

$$EuclideanDistance(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.2}$$

K-means uses an iterative refinement method. The final value is based on the user-defined number of clusters and the data set. Initially, K-means randomly chooses K as the mean values of K clusters, called centroids, and finds the nearest data points of the chosen centroids to form K clusters. Then, it iteratively recalculates the new centroids for each cluster until the algorithm converges to one optimum value [24].

3.2.2 K Value Selection

Elbow Method: is one of the most popular methods to determine this optimal value of K. The basic idea of the elbow rule is to use a square of the distance between the sample points in each cluster and the centroid of the cluster to give a series of K values. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The sum of squared errors (SSE) is used as a performance indicator. Iterate over the K-value and calculate the SSE. Smaller values indicate that each cluster is more convergent [28].

3.3 Deep Neural Network

3.3.1 Introduction

Artificial intelligence (AI) is any computer program that does something smart. It can be a stack of a complex statistical model or if-then statements. **Machine learning** is a subset of AI. Machines take data and 'learn' for themselves. It is currently the most promising tool in the AI pool for businesses.Machine learning allows a system to learn to recognize patterns on its own and make predictions, contrary to hand-coding a software program with specific instructions to complete a task. **Deep learning** is a subset of machine learning which is more applicable for the data analysis with increased non-linearity and more complexity in feature extraction [29].

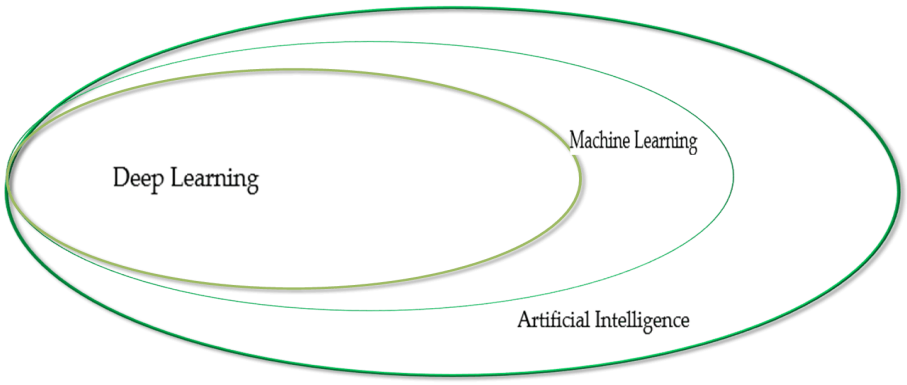

Figure 3.1: Artificial Intelligence, Machine learning, Deep learning [29]

3.3.2 Structure of Deep Neural Network

A neural network is composed of an input, an output and at least one hidden layer. The layers are composed of elements called neurons. The neurons of a layer are connected to neurons of the preceding and succeeding layers through weighted connections, also called network weights [30].

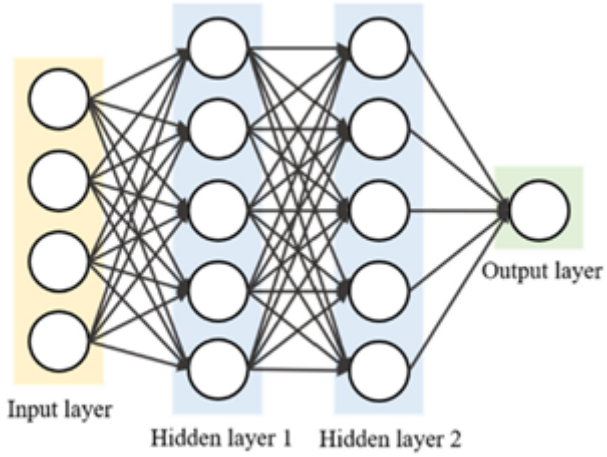

Figure 3.2: Artificial Neural Network Structure [30].

3.3.3 Deep Learning Models for Analysis of Spectrum Utilization

The Deep Learning algorithms used for this thesis are Long Short-Term Memory (LSTM) and Convolution Neural Network (CNN). A brief insight to the algorithms is as follows:

3.3.4 Long Short-Term Memory (LSTM)

Long short-term memory network is an advanced recurrent neural network (RNN). It is capable to learn order dependence in sequence prediction. The main limitation of RNN is unable to learn long-term dependencies due to vanishing gradient problem in the operation of backpropagation. LSTM handles vanishing gradient problems easily by using its four interacting layers within a cell which differs from standard RNN models as shown in Figure 3.3. The LSTM structure is built in the manner of a cell state that runs through the entire LSTM, the gates that work by either authorizing the value to be applied to the

cell state or altering the value by disallowing the data. There are also components called gated cells that allow the information to be stored in them from previous LSTM outputs or layer outputs [31]. The LSTM contains three parts, namely Foregate gate, Input gate

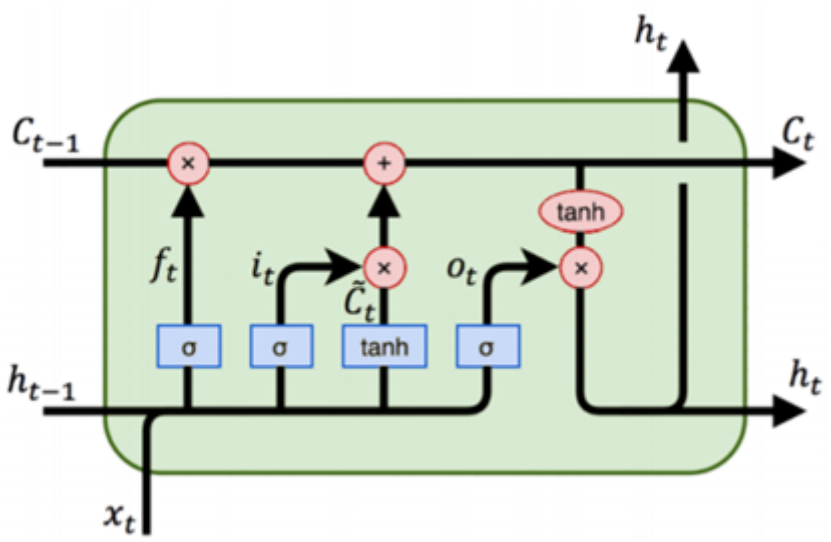

Figure 3.3: LSTM Architecture [32].

and Output gate and each part perform a separate function [33].

The Forget gate: chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. The sigmoid function is used to transmit information between the current input X_t and the hidden state H_{t-1}. The forget gate equation is as follows.

$$f_t = \sigma\left(X_t * U_f + H_{t-1} * W_f\right) \tag{3.3}$$

Where U_f is weight associated with the input and W_f is the weight matrix associated with hidden state.

Input gate is used to govern the significance of the new data carried by the input. First, a sigmoid function is used to combine the previous hidden state and the current input. By changing the values to be between 0 and 1, determines which values will be updated. Equation 3.4 describes the input gate equation.

$$i_t = \sigma\left(X_t * U_i + H_{t-1} * W_i\right) \tag{3.4}$$

Here i_t is input gate and U_i is weight matrix of input.

The new information is a function of a hidden state at the previous time stamp t-1 and input x at time stamp t. The activation function here is tanh. Due to the tanh function, the value of new information will be between-1 and 1. Cell state at the current time stamp (c_t) is as follows:

$$C_t = f_t * C_{t-1} + i_t \tan h\left(x_t * U_c + H_{t-1^*} W_c\right) \tag{3.5}$$

Output gate the cell passes the updated information from the current timestamp to the next timestamp.

$$O_t = \sigma\left(X_t * U_o + H_{t-1} * W_o\right) \tag{3.6}$$

The modified cell state's O_t and tanh determine the current hidden state. As indicated in the equation below.

$$H_t = O_t * \tan h\left(C_t\right) \tag{3.7}$$

The hidden state(H_t) turns out to be a function of the current output and long term memory (C_t).

3.3.5 Convolution Neural Network (CNN)

CNN is a type of deep neural network initially designed for image processing problems. However, it is now applied to data that can be represented in a grid-like matrix form. For instance, time-series and textual data can be represented by a 1D vector and a 2D matrix can be used to represent the pixels in the image data. The architecture was named "convolution neural networks" after the mathematical operation Convolution. It involves performing a linear operation in an ordinary matrix multiplication on at least one of the neural networks [34].

CNN has an input layer, hidden layers, and an output layer. The convolutional layer, pooling layer, and fully connected layer are primary processes used to form a CNN model for the goal of extracting features. The main purpose of a convolution layer is to detect and extract features from the input. Each convolution layer contains multiple convolution kernels (filters), which are used to convolve the input feature map so that it can generate the output feature map. The pooling layer is also known as the subsampling layer and is

applied to reduce the dimensionality of the feature map. It also reduces the number of parameters in the network to avoid overfitting. Finally, the fully connected layer flattens the result of a regular neural network and then produces the desired output, whether for classification or prediction [35].

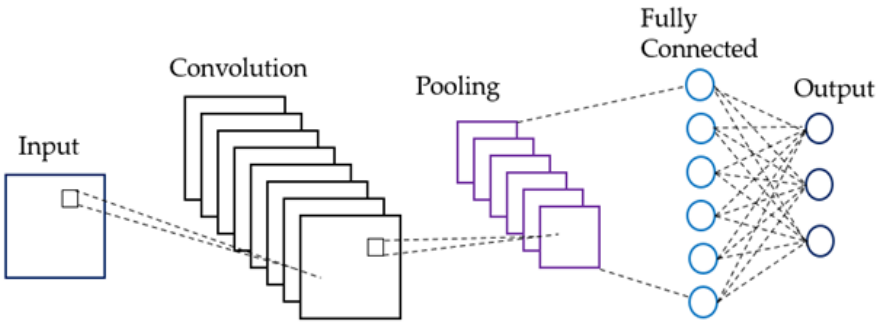

Figure 3.4: CNN Architecture [30].

CNN 1D convolutions to extract information along the time dimension. A convolution can be seen as applying and sliding a filter over the time series. Unlike images, the filters exhibit only one dimension (time) instead of two dimensions (width and height). The filter can also be seen as a generic non-linear transformation of a time series.

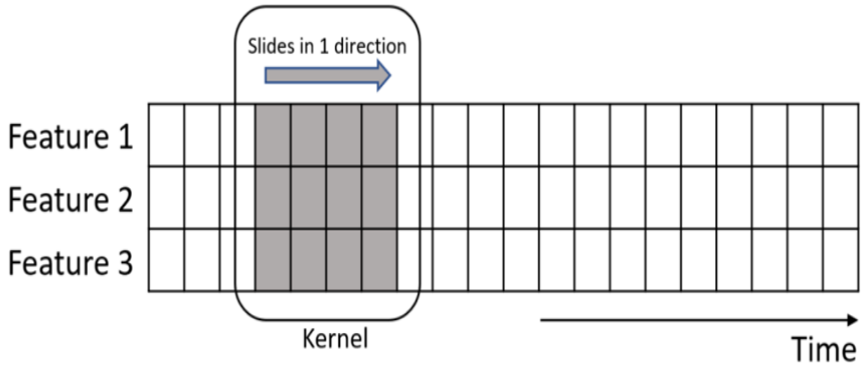

Figure 3.5: 1D CNN for Time Series Data [30].

3.4 Deep Neural Network Hyperparameters

3.4.1 Hyperparameters

Parameters are the coefficients of the model. They are chosen by the model itself. A hyperparameter, on the other hand, is a parameter that is set before the learning process

begins. These parameters are tunable and can directly affect how well a model train [36].Some hyperparameters are discussed below.

Number of Hidden Units and layers: Hidden units are the number of neurons in the hidden layer. Because the number of units in the hidden layer affects the model's performance, choosing the best number is crucial. In general, having too few hidden units leads to high training errors due to under-fitting. Too many hidden units will result in fewer training errors due to over-fitting. The size of hidden units depends upon the number of input training examples [37].

Activation Function determines the active state of the neuron. It decides whether the information received by the neuron is relevant or not. It transforms the input signal into the non-linear form and is sent to the next layer as an input [38].

Optimization is a process which tried to reduce the network error. This plays a crucial role in improving the accuracy of the model. The variants of optimizer are Stochastic Gradient Descent (SGD), Nesterov accelerated gradient, Adagrad, RMSProp, AdaDelta, Adam.

Learning Rate s used to tune the models. It minimizes the error by updating network weights. Choosing too low a learning rate and too high a learning rate would degrade model performance. The low learning rate will make tiny updates in the network weights and slow down the training process while too high a learning rate will cause divergent behavior in the error [36].

Number of epochs: The number of epochs represents the number of passes through the training dataset. Each epoch denotes that the training sample can change the model's internal parameters.

Batch size: Batch size is training samples passed to the networks. A training dataset can be divided into one or more batches. When all training samples are passed as a single batch, then the learning algorithm is known as batch gradient descent. When the batch size is one, then the learning algorithm is known as stochastic gradient descent. When the batch size is more than one and less than the training size, the learning algorithm is called mini-batch gradient descent. In mini-batch gradient descent, 32, 64, and 128 batch sizes

are more popular [38].

Dropout: is a regularization method. It involves removing some nodes so that the neural network. During training, some number of layer outputs are randomly dropped [36].

3.4.2 Hyperparameter Tuning

Hyperparameter tuning is finding the best hyperparameters to get the best results from your models. Hyperparameters are set before training a machine learning model. These hyperparameters need to be optimized to adapt a model to a dataset. Searching for the best hyper-parameter can be tedious, hence search algorithms like grid search and random search are used. Searching for the best hyper-parameter can be tedious, hence search algorithms like grid search and random search are used [38].

Random Search: In ransom search we create a grid of possible values for hyperparameters. Each iteration tries a random combination of hyperparameters from this grid. The performance is recorded and lastly returns the combination of hyperparameters which provided the best performance.

Grid Search: In grid search, we create a grid of possible values for hyperparameters. Each iteration tries a combination of hyperparameters in a specific order. Each potential combination of all of the hyperparameter values provided is built into a model, which is then evaluated before picking the parameters that produce the best result. However, with higher dimensional hyper-parameter space, grid search will inevitably suffer from the phenomenon known as curse dimensionality [39].

3.5 Model Performance Evaluation Metrics

The Model evaluation aims to estimate the generalization accuracy of a model on future or test data. Model evaluation metrics are used to quantify model performance. The evaluation is performed using standard prediction evaluation metrics [40].To evaluate the proposed method, four commonly used evaluation metrics are used to measure the efficiency of the models, the RMSE, MAE, MAPE, and MSE.

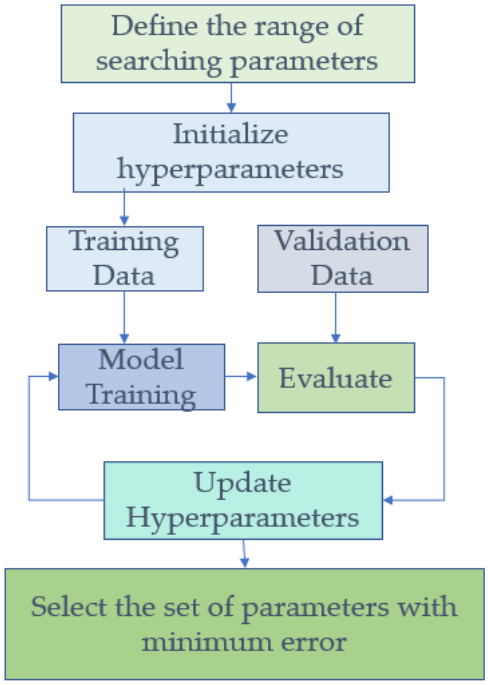

Figure 3.6: Hyperparameter Tuning .

Mean Squared Error (MSE): average squared difference between the predicted values and the actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.8}$$

Where y_i and$(\hat{y}_i)$ are actual and predicted values, respectively and N is the number of data points.

Mean Absolute Error (MAE): calculates the absolute difference between actual and predicted values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.9}$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.10}$$

Mean Absolute Percentage Error (MAPE): is similar to MAE only difference it we take the percentage error.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{3.11}$$

Chapter 4

Result and Discussion

4.1 System Model

The model building is expected to capture all the characteristics and components of the data set. Spectrum utilization data is modeled as time series data. The model is developed for a for the LSTM and CNN algorithms for clustered and non-clustered data.

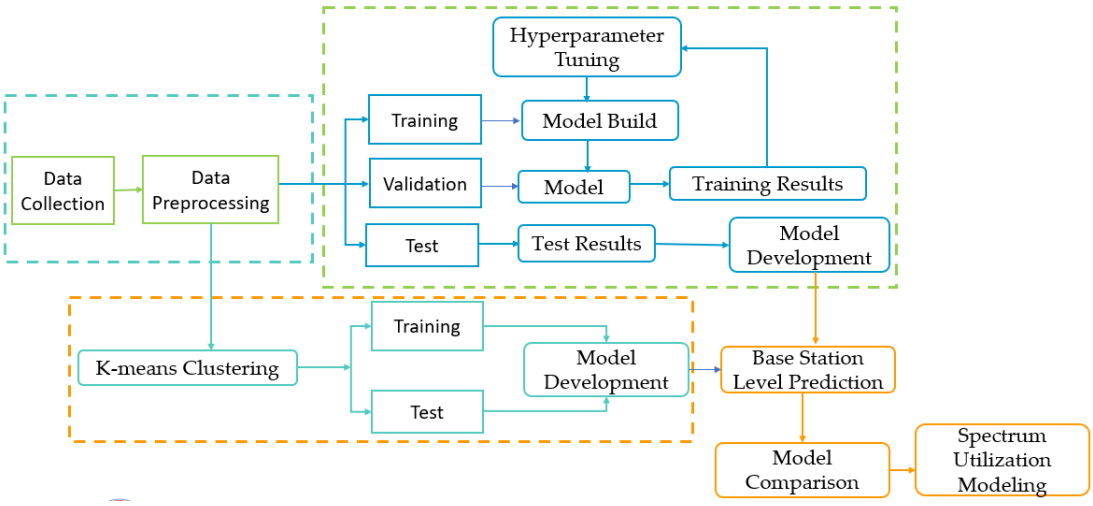

Figure 4.1: System Model

4.2 GSM Frequency Configuration in Ethio telecom

Each base station is given a percentage of the total number of channels available to the entire system. Nearby base stations are given distinct groups of channels to reduce interference between them. By systematically spacing base stations and their channel groups, the available channels are distributed throughout the geographic region and may be reused as many times as necessary, as long as the interference between co-channel of base stations is acceptable [11].

A traffic channel in GSM can carry user speech that can be either be in full-rate (TCH/F-13kbits/s) or half-rate (TCH/H-5.6kbits/s) channel mode [41]. When TCH/H is in use, one time slot may be shared by two connections. Thus, the efficiency of the spectrum in TCH/H will be doubled. Spectral efficiency can be measured in terms of users/cell of as more mobile users in a BTS is introduced to account for cellular coverage. In a GSM network, the configurations of users /cell depend on the physical hardware of the BTS. Typical configurations in GSM 900MHz are S222, S444, S666, and S888. It means, for example, S444, that there are 4 frequencies set in each sector of the antenna.

Ethio telecom has 61 and 85 channels configured to handle GSM service at 900Mhz and 1800Mhz, respectively. The maximum cell capacity for GSM900 is 8TRX per cell and 12 TRX for DCS 1800 [16].

Table 4.1: Ethio telecom GSM Channel Configration

Network Type	Channel Type	Frequency Number	Frequency Range
GSM900	BCCH	14	111-124
	TCH	47	63-109
	Gard frequency	1	110
GSM1800	BCCH	24	699-710,874-885
	TCH	61	712-743,844-872
	Gard frequency	2	711,873

4.3 Data Description and Pre-processing

4.3.1 Spectrum Utilization Data

In this thesis, spectrum utilization of GSM 900 in Addis Ababa, Ethiopia is studied. The data was collected for 100 days, from January 1 to April 10, 2021, from the Ethio telecom PRS, with a granularity of 1 hour for 639 sites. Additionally, the site's longitude and latitude are collected to analyze the spatial behavior of its utilization. Data preprocessing techniques such as handling missing values, standardization, and outlier handling, are applied to the collected dataset.

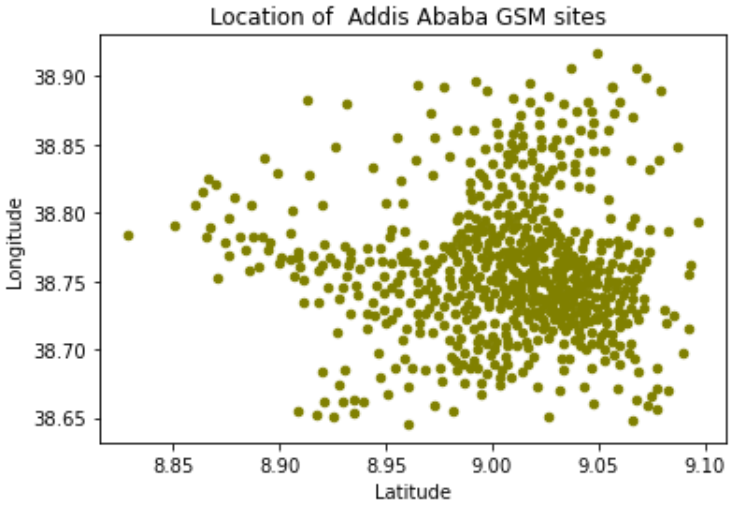

Figure 4.2: Base Stations Location

4.3.2 Data Pre-processing

The data set was gathered from 639 base stations (1917 cells). Five base stations with a continuous missing value were excluded. Some techniques were applied to replace the missing value. For instance, replacing the missing value with the next value, the previous value, and linear interpolation (the mean value of the previous and the next values). By computing the errors, the missing values were chosen to be replaced by linear interpolation.

The data set is divided into 80% test data, 10% validation data, and 10% test data for

model building, and open source libraries Keras and Tensorflow are used for implementation.

Table 4.2: Sample Records from the Dataset

Base Station	Date and Time	Spectrum Utilization	Latitude	Longitude
BTS1	1/1/2021 6:00	9.62%	38.72548	9.03244
BTS1	1/1/2021 10:00	43%	38.72548	9.03244
BTS2	1/1/2021 15:00	25.53%	38.90538817	9.03715001
BTS2	1/1/2021 21:00	97.99%	38.90538817	9.03715001

4.3.3 Dataset Decomposition

Time series have different components. Similarly, the spectrum utilization data set, as it is a time series, has daily and weekly seasonality, and it also has a nonlinear component, as shown in Figure 4.3.

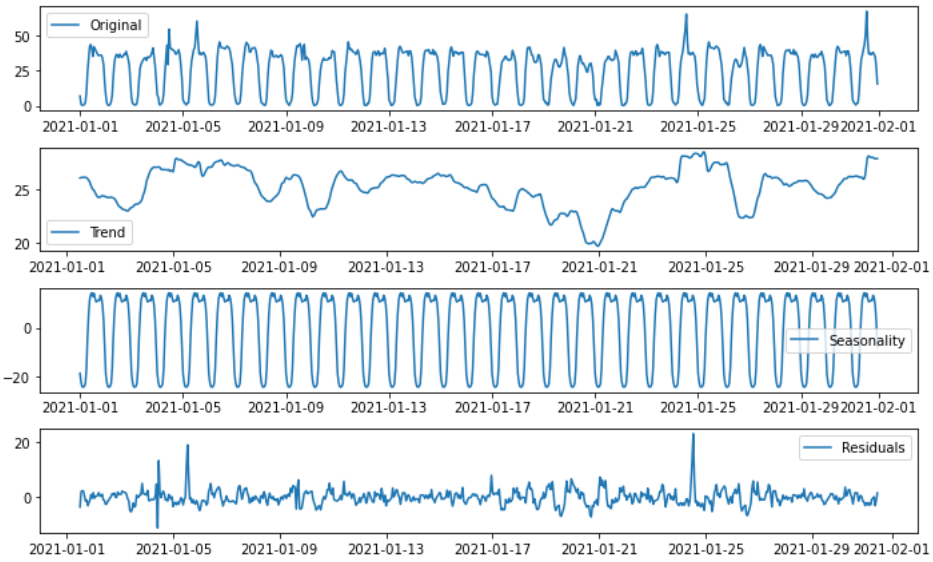

Figure 4.3: Time series Components Decomposition

4.4 Base Station Clustering Using K-means Algorithm

The cell load varies due to different factors. Changes in work or rest time in commercial and residential regions, as well as variations in human behavior through time and in different locations, are some factors. Figure 4.4 depicts the average spectrum utilization of four different base stations over the course of a week, with each base station's utilization levels and temporal usage characteristics differing from the others.

BTS may have significant load variation in hours of the day. Spectrum holes can be found in those periods when cells of one Radio Access Technology (RAT) have low load while the co-located cells of another RAT have the demand. This hole can be managed using load balancing while it needs farther investigation.

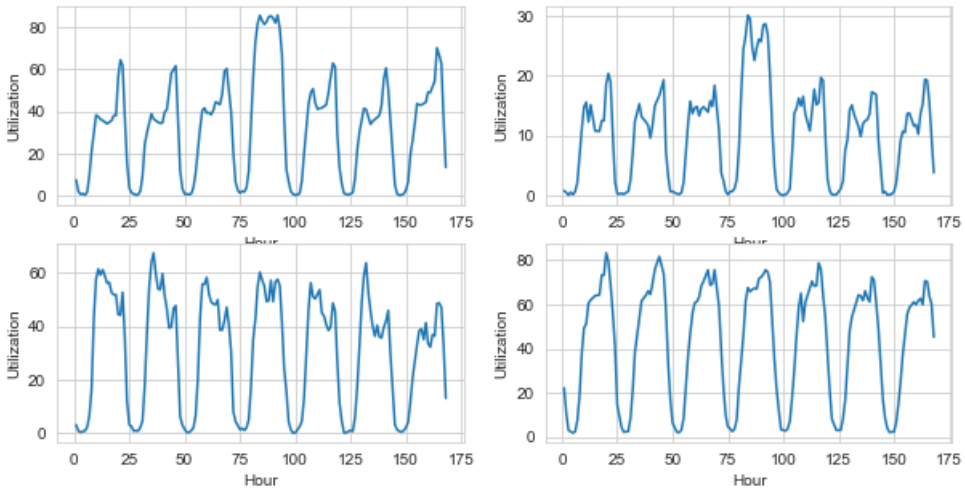

Figure 4.4: Different Channel Utilization at Base Stations

4.4.1 Clustering Based on Average Spectrum Utilization

Average spectrum utilization level clustering means that the average utilization level varies between different base stations, which means that some base stations may have high usage while others may have low utilization.

Figure4.5 shows the average spectrum utilization for 634 sites. As illustrated in the Figure, some BTSs have around 70% average utilization, while some have less than 5% average utilization. So that clustering can be used to group the BTSs with similar utilization levels.

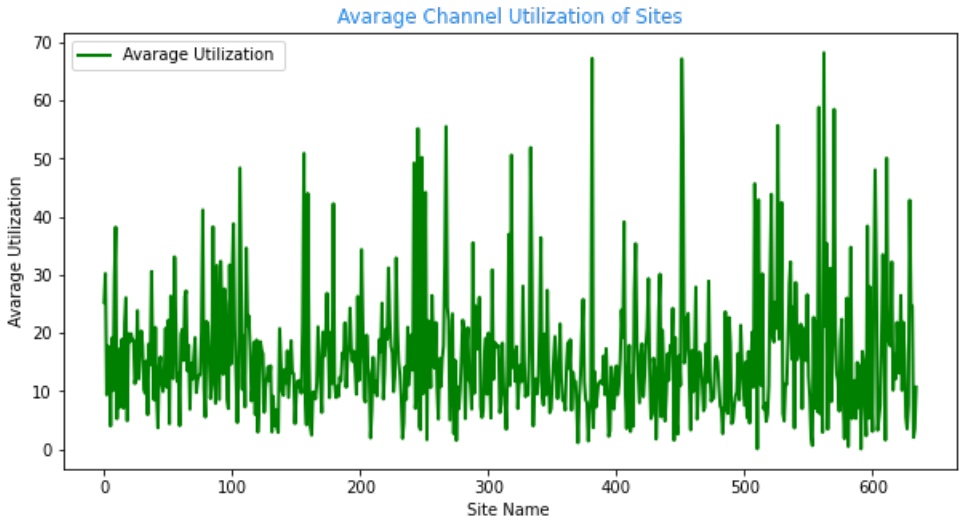

Figure 4.5: Average Channel Utilization

Optimal Number of Clusters, K

K-means algorithm requires a predefined k, the number of clusters. Using the elbow method dataset is grouped into three clusters. As shown in Figure 4.6 different locations have different utilization levels. In some areas and some BTSs may use their resources optimally, while others do not. Clustering base stations according to their utilization can be used to make radio resources more efficient. A load balancing mechanism, where overloaded BTSs of one RAT can distribute some of their traffic to less loaded BTS of other co-located RAT,can be implemented on BTS with low utilization level.

Scenario 1: High Utilization

According to the results, cluster 0's average utilization ranges from 35% to 70%. When attempting to analyze the sites, some sites have 100 percent utilization at peak hours, which may lead to call drops.

Some sites in this cluster 0 have high utilization due to the Inter-RAT Circuiting Switching (CS) service Handover Switch (CSServiceHOSwitch) being set to on. This switch is service-based UMTS-to GSM handover. When the switch is set to ON, the inter-RAT handover for CS services is enabled. The service-based inter-RAT handover supports UMTS-to-GSM handovers based on service attributes and can balance the load between the two systems by transferring some kinds of appropriate services to GSM/GPRS [42]. Turning on this switch results in UMTS-to-GSM handovers for most CS services and

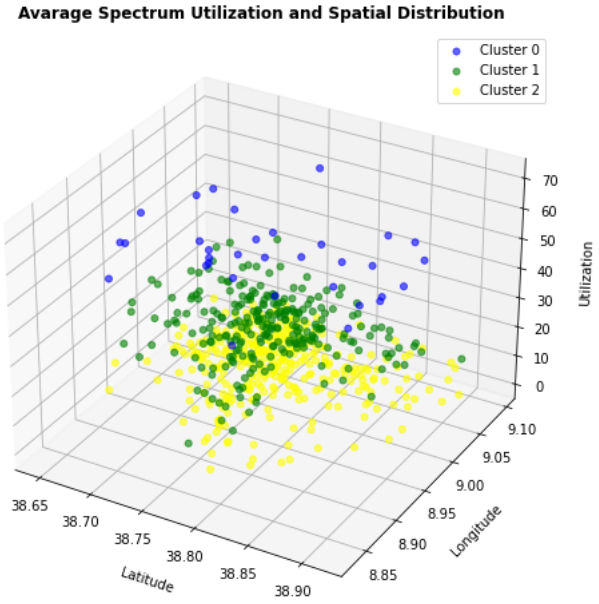

Figure 4.6: Clustering Result and their Spatial Distribution

reduces the UMTS resources occupied by voice in turn leads the GSM to be congested.

Another reason for high utilization is that the sites are GSM900 only sites. Because the half-rate configuration can double the channel capacity, it may be a viable option.

Scenario 2: Low Utilization

A significant number of cells in this study has low utilization. They may have assigned more TRX. In such cases capacity degrading, TRX degrading, can be deployed. Figure 4.6 depicts a sample site with low utilization its high usage was on holidays. When we look at this cell channel utilization, we see that it has a maximum utilization of 23.64%. During the 100 days, the maximum incoming traffic was 10.16erlang, and using this maximum load, 17 channels would be enough to handle it with the full rate configuration, but the site is configured with 33 channels. Here, at least one TRX can be dismantled and used on other sites, which will reduce unnecessary expenditure costs.

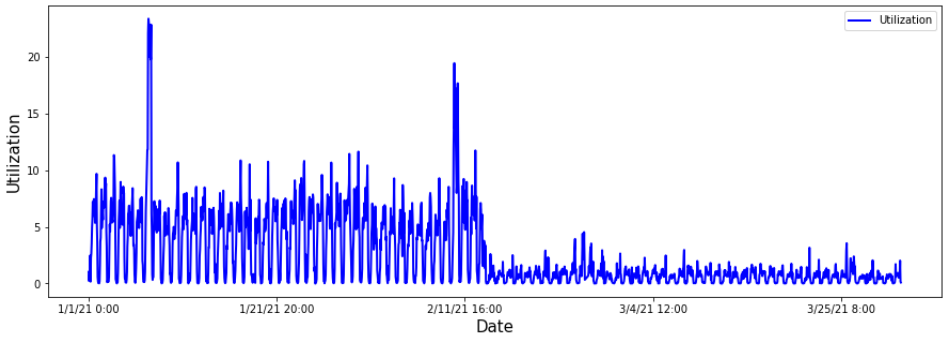

Figure 4.7: Sample Site Utilization

4.4.2 Clustering Based on Temporal Behavior of Spectrum Utilization

Spectrum utilization temporal behavior clustering means that utilization varies during different times. Groups of base stations may have high utilization during the day, while others may have high utilization during the night period. The amount of mobile traffic during the day or week depending on the living pattern of mobile users. Using preprocessed data, the K-means algorithm clusters the data into 5 clusters based on their temporal behavior. Figure 4.8 depicts the correlation of each cluster, and because clusters 1 and cluster 4 are highly correlated, four clusters are considered. .

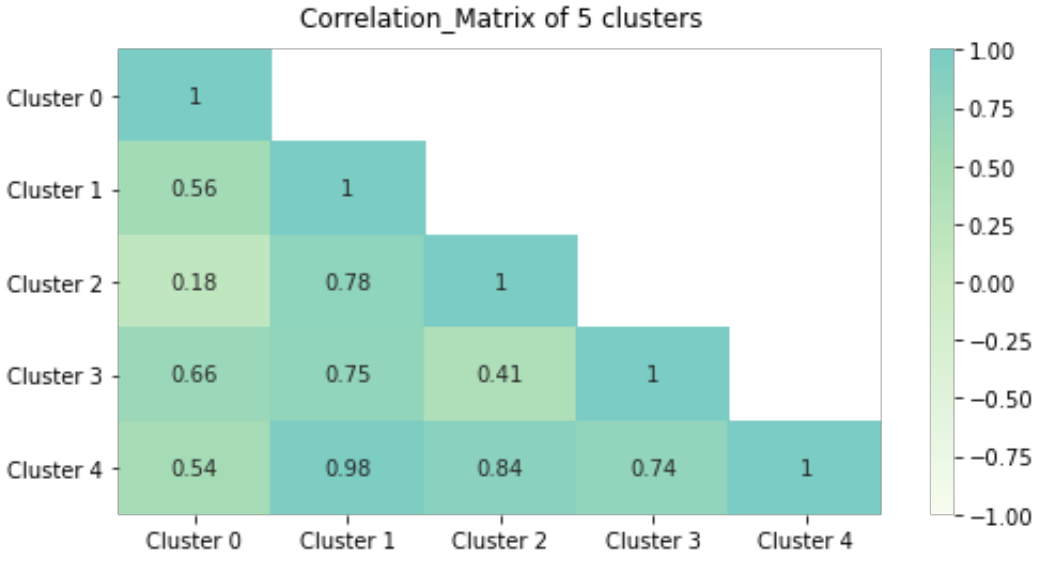

Figure 4.8: Correlation Matrix of Clusters

Figure 4.9 depicts a sample of the four clusters' daily spectrum utilization patterns. Cluster 0 base stations have high spectrum utilization during the day, while cluster 1 base

stations have high usage at night. Cluster 2 is heavily used throughout morning and night, while Cluster 3 is heavily used in the morning.

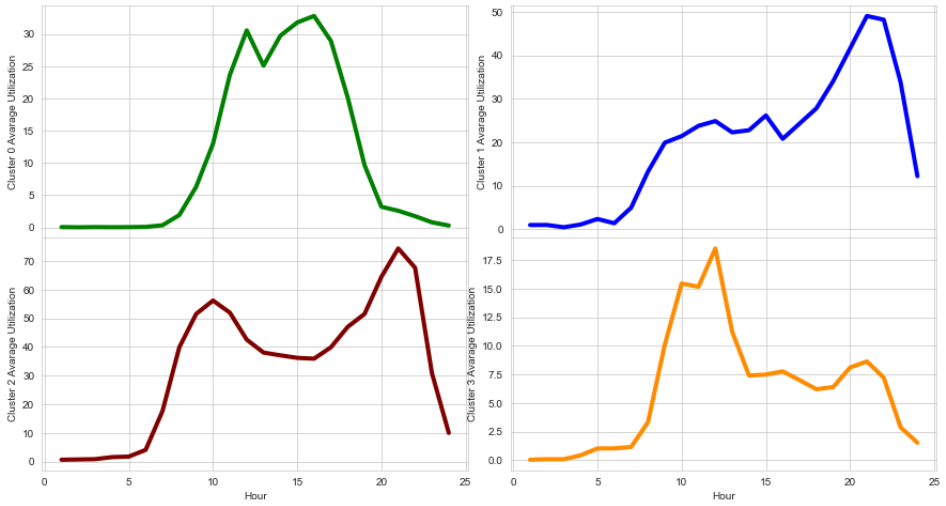

Figure 4.9: Daily Spectrum Utilization Pattern of Four Clusters

The following applications will result from modeling this variation using clustering, groups based on their similarity.

- It can be an input for better spectrum utilization through dynamic spectrum allocation, which is not yet implemented in Ethio telecom, but as it ends the monopoly, spectrum resources will be scarce resources, and different techniques will be implemented on multi-operator networks. The DSA can be between technologies, between operators or it can be in the context of cognitive radio.

This study result has shown that spectrum resources are not efficiently utilized. When we observe the spectrum overtime in a particular area, there exist spectrum holes. The in-efficient use of spectrum may result from fixed spectrum allocation (FSA) schemes. Operators can get economic benefits just by providing free resources. It can be providing a solution to the problem of in-efficient spectrum utilization by sharing spectrum dynamically between operators in a particular area under certain traffic conditions

- Clustering groups according to their utilization variation. Because the fluctuation in the same cluster is comparable, the cluster's off-peak hours will be similar. Due to this, it can be used to plan an energy-efficient network for Ethio telecom. During

off-peak hours by scheduling base stations off/on to reduce base station power consumption. Below, one sample site's utilization is considered to see how spectrum utilization analysis can be an input for energy efficiency network implementation.

Below one sample site utilization is considered to see how spectrum utilization analysis can be an input for energy efficiency network implementation.

Case Study

In the first cluster one site, which is located around Merkato, is used as a case study. BTSs of the four different cellular access technologies: GSM 900, GSM 1800, UMTS, and LTE Long Term Evolution (LTE) are located on the site. In terms of voice traffic, the chosen BTS is one of the busiest city sites. This site is used to show the relationship between BTS energy use and spectrum usage, or dynamic spectrum allocation and spectrum utilization. One week GSM900 and GSM 1800 channel utilization for the selected site is shown in Figure 4.10.

The largest energy consumer in the BTS is the power amplifier, which has a share of around 65 percent of the total energy consumption. The TRX is the most energy demanding part of the BTS hardware, and the number of TRXs inside the BTS cabinet influences the overall power consumption [43].The selected sample site GSM 1800 is configured with 12 transceivers (TRXs) per sector (12/12/12) and the GSM 900 configuration is 8 TRXs per sector (8/8/8). From the configuration, and as GSM 1800 is deployed for capacity utilization, there will be higher power consumption in the GSM 1800 BTS. But as we can see from Figure 4.10, during the night both BTSs have low utilization, so by considering this result, BTS sleep mode can be implemented. In the context of dynamic spectrum allocation, the sample site has low channel utilization during the night, so Ethio telecom can implement DSA with other operators..

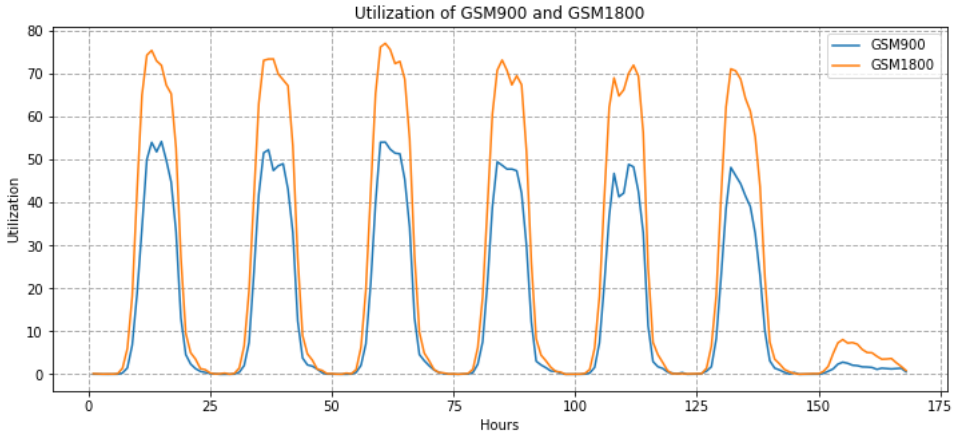

Figure 4.10: One Week Channel Utilization of GSM900 and GSM1800

4.5 Deep Learning-based Spectrum Utilization Modeling

The majority of time series forecasting algorithms require stationary time series, whereas deep learning does not. Deep learning is a subset of machine learning with more power and flexibility than in traditional machine learning approaches. The biggest advantage of deep learning algorithms is that they try to learn high-level features from data in an incremental manner. This prevents the need for manual feature extraction, which is done by human intervention [44].

The model building is expected to capture all the characteristics and components of the data set. Spectrum utilization data is modeled as time series data during model development. The model designed in this thesis divides the data set into training, validation and test data for the LSTM and CNN models. Starting in January 2021, the first 90% of the spectrum utilization data is used to capture the model's parameters and specifications for the training and validation set.As hyperparameters can affect the speed and accuracy of the final model, before final model deployment, the model is tested by validation date. The last 10% is used as the test data for performance comparison among the prediction models.

4.5.1 Spectrum Utilization Modeling Using LSTM

When building an LSTM model, it is required to consider how many hidden layers the model will include, the number of LSTM cells that should be used in each layer, and what the dropout should be. There is no right or wrong way to select the number of hidden layers or the number of cells within each layer or other hyperparameters. This number depends on the implementation of the LSTM model since the number of cells and layers will vary, but the layers are often from one to eight, and the cells in each layer can have the same number of cells to find the optimum structure [45].

A grid search algorithm was used to select LSTM model hyperparameters such as a hidden layer, batch size, and epoch. is used to determine the number of neurons in each layer. The settings that resulted in the least root mean squared error was chosen.

Using the results found in the grid search algorithm, the model is built using the hyperparameters listed in Table 4.3.

Table 4.3: Hyperparameters Used in LSTM Model

Hyperparameters	**value**
Hidden Layer	3
Hidden layer Neurons	Layer 1 48 Layer 2 32 Layer 3 32
Bach Size	128
Dropout	0.2
Optimizer	Adam
Activation	(ReLU)
Epoch	100

The model will either over-fit or under-fit the data. Over-fitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. Under-fitting refers to a model that can neither model the training data nor generalizes to new data [38].

As shown in Figure 4.11 above, the training and the validation loss of the model are

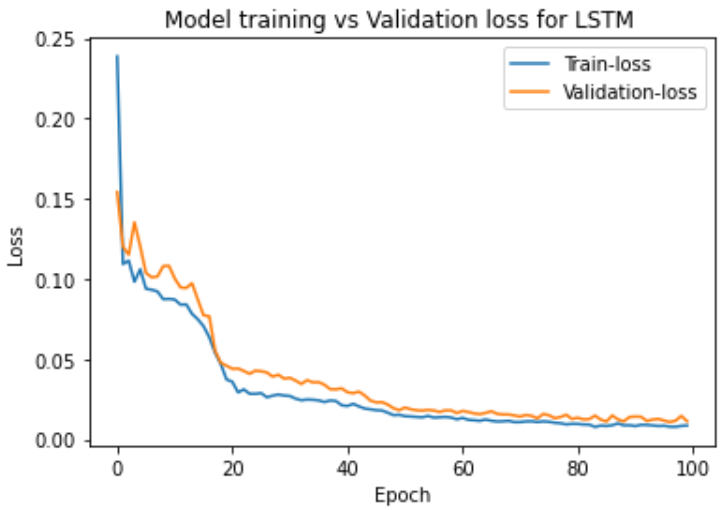

Figure 4.11: Plot for Training and Validation Loss of LSTM Model

plotted to see the possibility of overfitting and underfitting in our model. It is seen that there is no significant difference between the loss of validation and training.

4.5.2 Spectrum Utilization Modeling Using CNN

Hyperparameters are external parameters. Its values are used to control the learning process. The CNN model is tested with various hyperparameters such as kernel size, filter size, hidden layer, optimizer activation function, and Epoch. The Kernel size determines the height and width of the convolution window. Increasing kernel size helps to view more information inside set grids. Selecting these appropriate combinations of hyperparameters is critical to building a model with better accuracy. The grid search approach is used to determine the optimum hyperparameters by considering a best grid of possible values for hyperparameters, as shown in Table 4.4.

The model is built using the selected hyperparameters. Figure 4.12 illustrates the training and validation loss results.

The developed model is tested using the test data, which is 10% of the total data, or 10 days. Figure 4.13 shows the actual and modeled version graphs for the test data using LSTM and CNN algorithm.

Table 4.4: Hyperparameters Used in CNN Model

Hyperparameters	value
Number of Filters	64
Kernel Size	[3*3]
Bach Size	64
Hidden Layer	50
Dropout	0.2
Optimizer	Adam
Activation	(ReLU)
Epoch	60

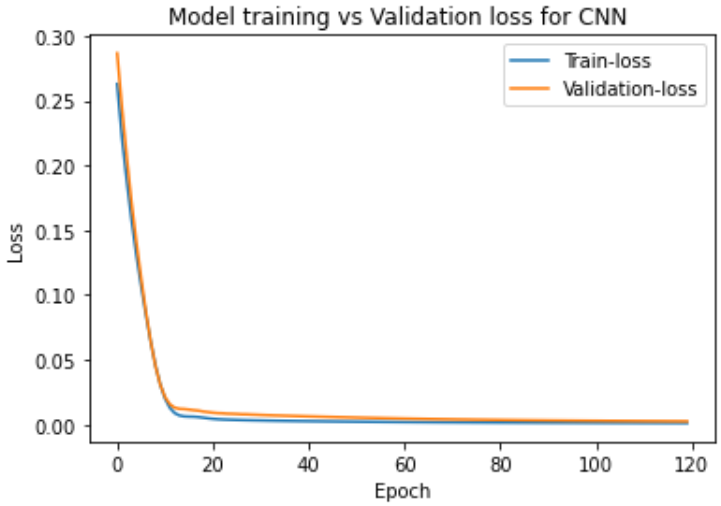

Figure 4.12: Plot for Training and Validation Loss of CNN Model

4.5.3 Spectrum Utilization Modeling for Clustered Data

The larger the training data set, the better the performance in time series forecasting. But, in the case of the data encountering different behaviors, this may not improve the performance [31]. Hence, the base stations that have similar traffic loads need to be trained together. To solve this, clustering was used in this thesis to group the base stations with similar spectrum utilization levels and spectrum utilization temporal behavior together. The reason for clustering the input is to achieve a high prediction accuracy of the customized model with a few inputs' parameters.

BTSs have different spectrum usage and, according to their usage in Section 4.3.2, there

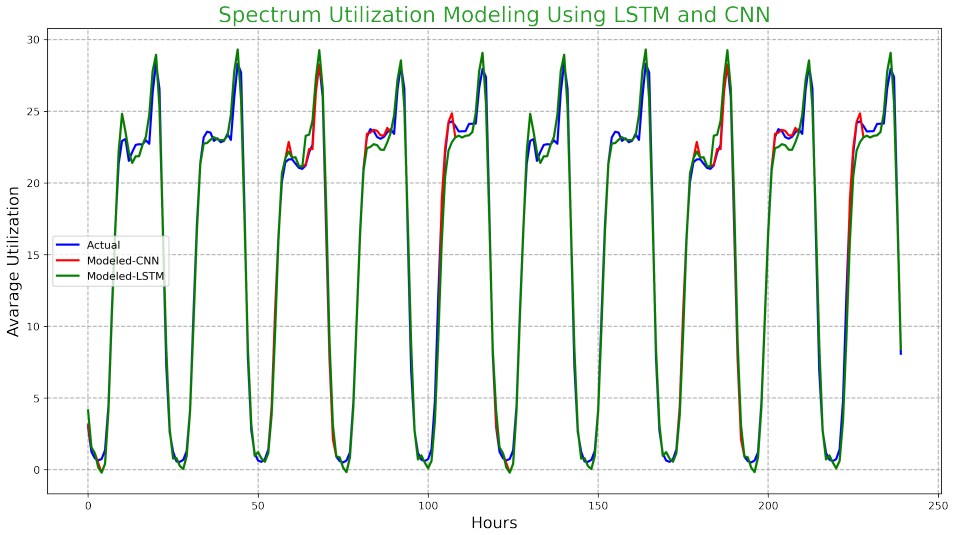

Figure 4.13: Actual vs. Modeled Plot for CNN and LSTM

are four clusters found in the Addis Ababa GSM 900 network spectrum usage. One of the clustered data from the four obtained cluster models is used to build a model using LSTM and CNN. To build this model hyperparameters used in Table 4.3 and Table 4.4 are used.

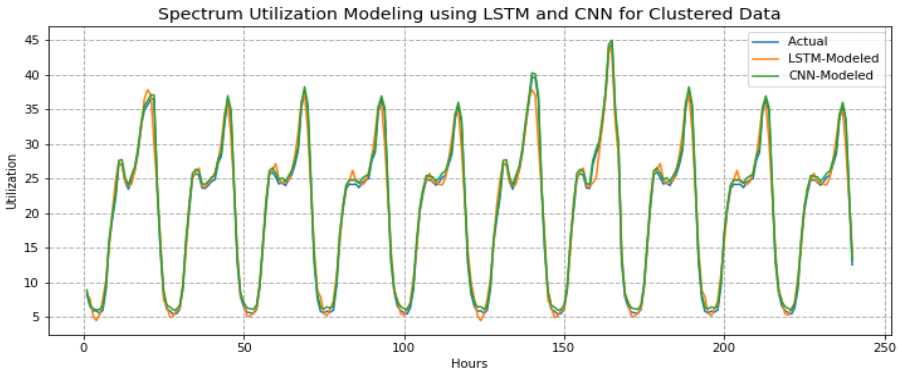

Figure 4.14: Plot of Actual Cluster Data vs. Modeled version.

4.6 Base Station Level Prediction

Base station level prediction is done for one site for twenty-four hours and the prediction is done using developed models.

As shown in the Table 4.7 below, the model developed by CNN algorithm using clustered data can predict up to twenty-four hours of future data traffic with an RMSE of 1.04.

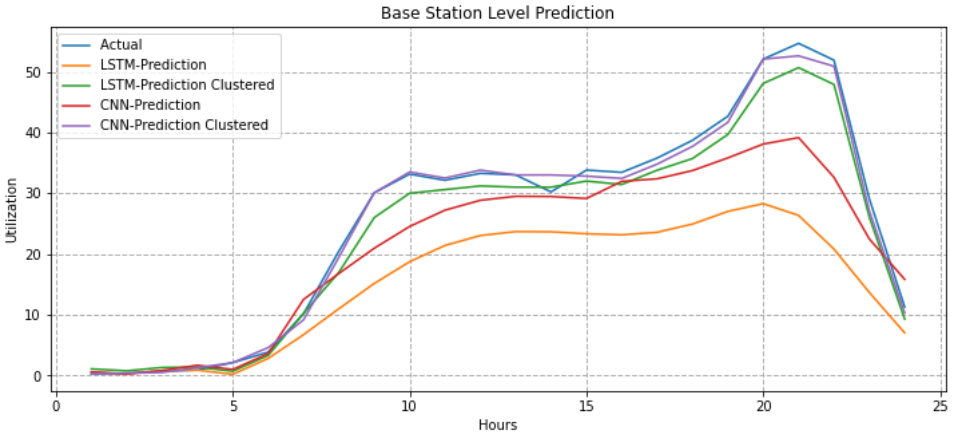

Figure 4.15: Base Station Level Prediction

4.7 Model Performance Comparison

Model evaluation aims to estimate the generalization accuracy of a model on test (unseen) data. Four basic prediction metrics, MSE, MAPE, RMSE, and MAE, are used to evaluate the prediction performance of a developed model. The prediction's accuracy is determined by the estimation of error; consequently, the lower the MAE, RMSE, and MAPE values, the better the forecast.

Table 4.5: Performance Evaluation Result

	Clustered				Non-clustered			
	MSE	RMSE	MAE	MAPE	MSE	RMSE	MAE	MAPE
LSTM	0.641	0.8	0.845	6.872	1.434	1.197	1.057	7.684
CNN	0.201	0.58	0.26	2.381	0.589	0.767	0.521	4.433

The values of MAPE and their interpretation, which are shown in Table 4.6 are used for the explanation of the prediction accuracy.

Table 4.6: MAPE Value Interpretation [46]

MAPE	**Interpretation**
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

The model developed by CNN using cluster data has the least errors of all of the proposed

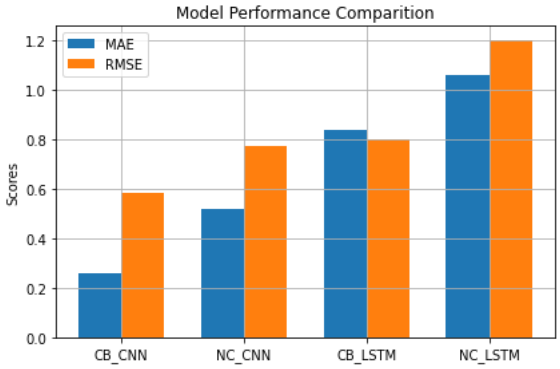

Figure 4.16: Model Performance Comparison

models.The CNN algorithm has also better computational cost.

4.7.1 Models Performance on Base Station Level Prediction

Base station prediction performance results are compared by RMSE and MAE, as shown in Table 4.7. The CNN model developed using clustered data outperforms the others.

Table 4.7: Performance Evaluation Result for Base station Level Prediction

	RMSE	MAE
LSTM-Prediction	13.34	10.325
LSTM-Prediction Clustered	2.4492	2.092
CNN-Prediction	7.114	5.058
CNN-Prediction Clustered	1.04	0.76

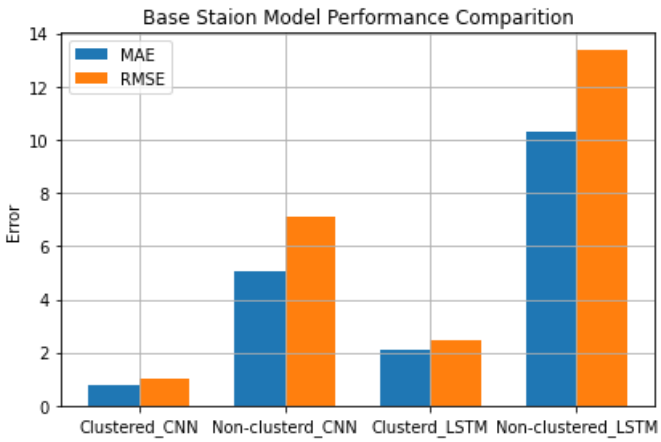

Figure 4.17: Base Station Model Performance Comparison

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

Using data from the Ethio telecom PRS, Addis Ababa GSM 900 network, spectrum utilization analysis and modeling are performed. The method used in this thesis can be applied to GSM 1800 as well as different geographical areas of Ethio telecom.

The result provides the opportunity for operators to know the extent to which the spectrum is utilized and its variation over time and space. Different clusters were identified based on the study, which can be used as an input for the Ethio telecom optimization techniques implementation. Spectrum utilization data is modeled as time series data. During model development, various strategies for enhancing the model's performance are employed to obtain a better model with the least amount of error. The model is developed using CNN and LSTM algorithms for clustered and non-clustered data. Comparison and evaluation are done using RMSE, MSE, MAE, and MAPE. The model developed for the cluster data using the CNN algorithm can model spectrum utilization with RMSE value of 0.58 and this model after a twenty-four-hour observation of historical data it can predict next twenty-four –hour base station spectrum utilization with an RMSE value of 1.04.

5.2 Recommendation

The study of spectrum utilization using clustering and deep learning model result will be an input for the following recommendations:

- Currently, Ethio telecom LTE subscribers are increasing, and UMTS is getting some relief. So, UMTS can handle voice traffic. Thus, proper UMTS-to-GSM handovers optimization parameter setting can minimize network congestion due to high spectrum utilization.
- TRX degrading can be performed in sites with low utilization, avoiding unnecessary expenditure cost. So, this model will help to identify sites with low utilization.
- According to the result all sites do not use their spectrum optimally, which provides opportunities for implementing dynamic spectrum assess between operators between technologies and in context of cognitive radio.
- The result can be an input to implement cellular network strategies to reduce BSs energy consumption by scheduling switched off base stations.

According to the results, the spectrum is not utilized in an optimal manner, so as future work to improve spectrum utilization, apply dynamic spectrum access and study the impact on spectrum efficiency improvement and its techno-economic effect. In addition to have clear picture on Ethio telecom spectrum usage analysis for 3G and 4G spectrum usage can be done as future work.

Reference

[1] M. Cave *et al.*, "Review of radio spectrum management," *An independent review for Department of Trade and Industry and HM Treasury (www. spectrumreview. radio. gov. uk)*, 2002.

[2] R. Hafez and G. Chan, "Measures of the spectrum utilization," in *VTC'98. 48th IEEE Vehicular Technology Conference. Pathway to Global Wireless Revolution (Cat. No. 98CH36151)*, IEEE, vol. 1, 1998, pp. 277–281.

[3] S. Series, "Spectrum occupancy measurements and evaluation," 2017.

[4] V. Valenta, R. Maršálek, G. Baudoin, M. Villegas, M. Suarez, and F. Robert, "Survey on spectrum utilization in europe: Measurements, analyses and observations," in *2010 Proceedings of the fifth international conference on cognitive radio oriented wireless networks and communications*, IEEE, 2010, pp. 1–5.

[5] L. Mendes, L. Gonçalves, and A. Gameiro, "Gsm downlink spectrum occupancy modeling," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, IEEE, 2011, pp. 546–550.

[6] A. Agarwal, A. S. Sengar, and R. Gangopadhyay, "Spectrum occupancy prediction for realistic traffic scenarios: Time series versus learning-based models," *Journal of Communications and Information Networks*, vol. 3, no. 2, pp. 35–42, 2018.

[7] S. D. Barnes, P. J. Van Vuuren, and B. Maharaj, "Spectrum occupancy investigation: Measurements in south africa," *Measurement*, vol. 46, no. 9, pp. 3098–3112, 2013.

[8] L. Pedraza, C. Hernandez, and E. Rodriguez, "Modeling of gsm spectrum based on seasonal arima model," in *Proceedings of the 6th IEEE Latin-American Conference on Communications, Cartagena, Colombia*, 2014, pp. 5–7.

[9] M. S. Saitwal, "Dynamic spectrum sharing between mobile network operators in gsm," Ph.D. dissertation, Indian Institute of Technology Hyderabad, 2015.

[10] J. A. Musey and B. Keener, *The Spectrum Handbook 2018*. 2018.

[11] I. Marsa-Maestre, T. Ito, S. Pollin, A. Chiumento, and J. M. Gimenez-Guzman, *Efficient spectrum usage for wireless communications*, 2019.

[12] P. Dalela, A. Nayak, V. Tyagi, and K. Sridhara, "Analysis of spectrum utilization for existing cellular technologies in context to cognitive radio," in *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, IEEE, 2011, pp. 585–588.

[13] V. Valenta, Z. Fedra, R. Maršálek, G. Baudoin, and M. Villegas, "Analysis of spectrum utilization in suburb environment—evaluation of potentials for cognitive radio," in *2009 International Conference on Ultra Modern Telecommunications & Workshops*, IEEE, 2009, pp. 1–6.

[14] C. Srinuan and E. Bohlin, "A country comparative study of spectrum re-farming: Implication for thailand," 2018.

[15] M. SIDDIQUE, H. NAWAZ, and H. SOOMRO, "Comparative analysis and propagation modeling of umts 900/2100 mhz using matlab and signal pro simulations," *Sindh University Research Journal-SURJ (Science Series)*, vol. 45, no. 3, 2013.

[16] E. telecom, "Ethio telecom planning document," 2019.

[17] J. Eberspächer, H.-J. Vögel, C. Bettstetter, and C. Hartmann, *GSM-architecture, protocols and services*. John Wiley & Sons, 2008.

[18] K. Ikkelä, M. Myllynen, J. Heinanen, and O. Martikainen, "4g mobile network architecture," in *Emerging Personal Wireless Communications*, Springer, 2002, pp. 183–195.

[19] P. Stuckmann, "Traffic engineering concepts for cellular packet radio networks with quality of service support," Ph.D. dissertation, Bibliothek der RWTH Aachen, 2003.

[20] T. T. Nielsen and J. Wigard, *Performance enhancements in a frequency hopping GSM network*. Springer Science & Business Media, 2000.

[21] P. Stuckmann, "Traffic engineering concepts for cellular packet radio networks with quality of service support," Ph.D. dissertation, Bibliothek der RWTH Aachen, 2003.

[22] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*. prentice hall PTR New Jersey, 1996, vol. 2.

[23] S. Green, "Time series analysis of stock prices using the box-jenkins approach," 2011.

[24] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1361, 2019, p. 012 015.

[25] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster analysis 5th ed*, 2011.

[26] N. Chen, T. Qiu, X. Zhou, K. Li, and M. Atiquzzaman, "An intelligent robust networking mechanism for the internet of things," *IEEE Communications Magazine*, vol. 57, no. 11, pp. 91–95, 2019.

[27] K. Manjang, "Identification of customer profiles from electricity consumption data," 2018.

[28] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 336, 2018, p. 012 017.

[29] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071–1092, 2020.

[30] O. Ishaq, "Image analysis and deep learning for applications in microscopy," Ph.D. dissertation, Acta Universitatis Upsaliensis, 2016.

[31] T. A. Rashid, P. Fattah, and D. K. Awla, "Using accuracy measure for improving the training of lstm with metaheuristic algorithms," *Procedia Computer Science*, vol. 140, pp. 324–333, 2018.

[32] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[33] J. Moolayil, J. Moolayil, and S. John, *Learn Keras for Deep Neural Networks*. Springer, 2019.

[34] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, Ieee, 2017, pp. 1–6.

[35] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications surveys & tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[36] V. Alto, "Neural networks: Parameters, hyperparameters and optimization strategies," *Medium: Towards Data Science*, 2019.

[37] N. Mahamad and M. K. M. Amin, "Effect of number of hidden neuron in multilayer perceptron for power prediction," in *JSST 2013: Proceedings of the International Conference on Simulation Technology*, 2013.

[38] U. Michelucci, *Applied Deep Learning*. Springer, 2018.

[39] J. Jordan, "Hyperparameter tuning for machine learning models," *Retrieved from: Jeremy Jordan: https://www. jeremyjordan. me/hyperparameter-tuning*, 2017.

[40] M. R. Islam, N. A. Turzo, and P. S. Bishal, "Prediction analysis of gaming cost by employing data mining algorithms," 2021.

[41] K. Sudhindra and V. Sridhar, "An overview of congestion relief methodologies in gsm network," in *2011 7th International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, 2011, pp. 1–4.

[42] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 91–100, 2016.

[43] T. Chen, H. Kim, and Y. Yang, "Energy efficiency metrics for green wireless communications," in *2010 International Conference on Wireless Communications & Signal Processing (WCSP)*, IEEE, 2010, pp. 1–6.

[44] S. A. Graham, E. E. Lee, D. V. Jeste, *et al.*, "Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review," *Psychiatry research*, vol. 284, p. 112 732, 2020.

[45] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[46] C. Lewis, "International and business forecasting methods butterworths: London," 1982.

Appendix

Hybrid K-means Clustering and Deep Learning-based Spectrum Utilization Model in Spatial and Temporal Domains: A case study in the 900 MHz range

*Note: Sub-titles are not captured in Xplore and should not be used

1st Frehiwot Bantigegn
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
fruit.blove@gmail.com

2nd Dereje Hailemariam
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
dereje.hailemariam@aait.edu.et

***Abstract*—Abstract— Radio spectrum is a finite resource while the demand for wireless systems is increasing at an exponential rate. To meet this demand, new generations of cellular networks were introduced. Spectrum utilization of cellular bands is analyzed widely using spectrum measurements. Knowledge of the spectrum utilization will help operators like, Ethio telecom to understand and plan band usage. In this paper, using the K-means algorithm and Deep learning algorithms, namely Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), downlink Global System for Mobile Communication (GSM) 900 spectrum utilization is analyzed and modeled, and base station level prediction is done. The data were collected from Addis Ababa 639 base stations. Spectrum utilization is modeled using CNN and LSTM algorithms for clustered and non-clustered data. Because of the differences in base station spectrum utilization, it is better to cluster the base stations before model development. The CNN model trained on clustered data outperforms the other models, mainly for base station level prediction. Our results show that the GSM 900 downlink spectrum is not utilized optimally. The highest observed average spectrum utilization was 71%, with the lowest observed average spectrum utilization being 1.4%.**

***Index Terms*—Spectrum Utilization, GSM900, K-means, LSTM, CNN**

I. Introduction

The tremendous growth of wireless services has created an ever-increasing demand for the radio frequency spectrum. Radio spectrum, however, is a finite resource and has been intensively used so far. To support this growth in demand, operators and regulators are examining ways to 1increase the available spectrum and efficient utilization of the existing spectrum [1].

The term "spectrum utilization" refers to the amount of information (measured in bits) that is being carried by a spectrum unit (measured in m2. Hz. Sec.). Radio spectrum is defined along three dimensions: space, frequency, and time. Parameters that limit the full utilization of the available radio spectrum is the non-uniformity of user density [2].

Static frequency allocation leads to vast underutilization of frequency spectrum due to random usage within various geographical regions and time [3].

Despite all the attempts, the spectrum crunch will continue to surface forcing many wireless network operators to look towards other solutions such as dynamic spectrum access with cognitive capability, refarming their 2G spectrum, and deploying more spectrally efficient technologies like Long Term Evolution (LTE) and Fifth generation(5G) networks. With the current spectrum management policy, which pre-allocates the spectrums for authorized users, some bands have heavy traffic on them, while the other bands are left idle most of the time, leading to a huge waste of spectrum resources [4] [5].

To increase spectrum usage efficiency, it becomes necessary to have a clear picture of how different frequency bands are used in specific environments.

The network dimensioning is usually performed based on the estimated load in typical busy-hours conditions. As a result, the allocated carriers are well utilized during peak hours, e.g., daytime in business areas, offices, they become underutilized for the rest of the time, e.g., nighttime. Moreover, the traffic demand may increase in other areas of the network, e.g., in residential areas in the evening, located outside the busy hour high traffic areas. Operators are expected to continuously monitor the utilization of their spectral resource, which is the availability of spectrum in terms of space (e.g., location, service area), time, and a number of channels (in a channelized band) that all users in a certain territory may access. Spectrum measurements can be used to assess the current status of the spectrum use and availability of the spectrum for other users.

Operators shall understand the dynamics of spectrum utilization assessment of frequency occupancy (current and future) in both spatial and temporal domains and reallocate

Identify applicable funding agency here. If none, delete this.

spectrum resources according to the spatial and temporal traffic requirements. Hence, the main motivation of this paper is, to understand (in an average sense) how the operator utilizes the different spectrum bands allocated to it, in both spatial and temporal dimensions. For that traffic channel utilization reports (hourly based frequency utilization per cell in percentage) generated from 600++ base stations of the operator network were collected for 100 days. Measurements were taken on an hourly basis for the GSM frequency range. Using data collected from traffic channel utilization less necessitates using external spectrum analyzer. Measurements are taken for 24 hours in 600++ locations for 100 days, which makes the use of a spectrum analyzer less necessary. Moreover, spectrum measurement using a spectrum analyzer is costly. It covers a limited geographic area and limited time slots of the day. With knowledge of its spectrum utilization, the company may follow multiple approaches (going for a new frequency band in the case of "full utilization"; reframing frequency, half rate configuration implementing spectral efficient technologies, ... to improve the utilization; or in the case of low utilization even allowing other users to utilize its frequency in the context of cognitive radios or spectrum sharing with other operators.

The Benefit of an indirect occupancy assessment – low cost and less resource as it uses the already available data in an operator.

A. Literature Review

Most of the spectrum utilization assessments reviewed in the work are conducted to explore the possibility of employing dynamic spectrum access in the context of cognitive radio.

Motivated by the lack of knowledge regarding spectrum occupancy in South Africa, the authors in [7] measured the spectrum occupancy in the ultra-high frequency (UHF), GSM 900 MHz ,and 1800 MHz bands. Measurements were taken on a daily basis, spaced at two hourly intervals for six weeks, and with a sampling resolution of 2Mhz for UHF and 100 kHz for the GSM bands. The results indicate a maximum occupancy of 20% for UHF bands. For the GSM bands, during peak hours the maximum utilization for the GSM 900 MHZ and 1800 MHz bands are 92% and 40%, respectively .

In [8] GSM channel utilization modeling and prediction are done. Spectrum measurements were performed in Bogota City, Colombia, in the GSM band of 850MHz.For 7 days, 60 channels were measured in this band. To model and predict the channel utilization Seasonal Autoregressive Integrated Moving Average (SRIMA) has been used. The result of primary users (PUs) occupancy models can be used as empty channel indicators replacing the spectrum sensing procedures. Besides that, based on prediction information, SUs can select the channels with a higher probability of availability in multi-channel wideband sensing scenarios.

Based on a spectrum measurement in Jaipur, Rajasthan, India, the study in [4] analyzes the practical prowess of time-series modeling methodologies, namely, Autoregressive (AR) and Autoregressive Integrative Moving Average (ARIMA) models, and machine learning techniques, namely, Lagrangian Support Vector Machine (LSVM) and simplified models of RNNs, i.e., Elman network (EN) for predicting spectrum occupancy in a TV band, and in a cellular band. for predicting spectrum occupancy in a TV band and in a cellular band. The measurements were performed over one week over the frequency range of 150-750 MHz in the TV band, and 850-1 300 MHz in the cellular band. The performance evaluation is done using MSE and the study discovers that due to cellular data traffic is non-stationary with several irregularities, the RNN technique outperforms the other model in terms of prediction accuracy. While in the TV band, the traffic pattern is stationary, and the time-series models can work efficiently.

Paper [5] addresses the problem of inefficient spectrum utilization in GSM using dynamic spectrum sharing (DSS) between mobile network operators. The proposed spectrum sharing scheme considers spectrum utilization and call blocking probability and is evaluated under the different traffic conditions for base stations. The result for the proposed scheme shows improvement in spectrum utilization with reduced call blocking probability. From the above reviews, we have seen that:

- Whether to add other services or to know how effectively utilize spectrum operators should know to what about they are using their spectrum.
- Static frequency allocation leads to vast underutilization of frequency spectrum due to random usage within various geographical regions and time.
- Due to their capacity to capture the spectrum utilization dataset's non-stationary behavior deep learning models are recommended for spectrum usage prediction over time series prediction methods.

The main objective of this paper is to develop a model that captures the space and time variation of spectrum utilization. With the above intentions, the approach followed in this work are the following:

- For the 600++ base stations, apply K-mean clustering to classify the utilization levels into "High", "Medium" and "low" to see the overall utilization level
- For the 600++ base stations, apply K-mean clustering to group the same temporal behavior of spectrum usage and will be used for optimization purposes like for resource allocation, for implementing dynamic spectrum access and in addition to this it can improve the prediction accuracy
- Apply LSTM and Convolutional Neural Network to predict future utilization per cluster level.

To the best of our knowledge, no prior work is done to investigate spectrum utilization based on the operator's data. For the case of GSM1800 networks, the same approach can be extended.

II. SPECTRUM UTILIZATION ANALYSIS

A. Spectrum Utilization Dataset

In this paper, spectrum utilization of GSM 900 in Addis Ababa, Ethiopia is studied. The data is obtained from the Ethio

telecom performance resource system (PRS) for 100 days from January 1 to April 10, 2021, with a granularity of 1hr for 639 sites. Additionally, the site's longitude and latitude are collected to analyze the spatial behavior of its utilization.

As shown in Fig1 there is significant channel utilization variation on the GSM network.

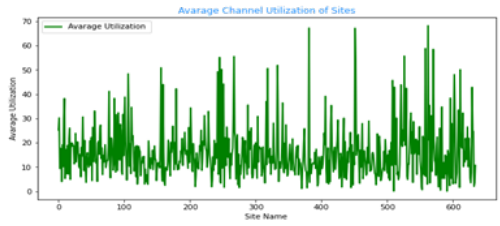

Fig. 1. Average Channel Utilization.

B. Methodology

Spectrum utilization data is modeled as time series data. The model is developed using LSTM and CNN for clustered and non-clustered data. The overall methodology is as follows:

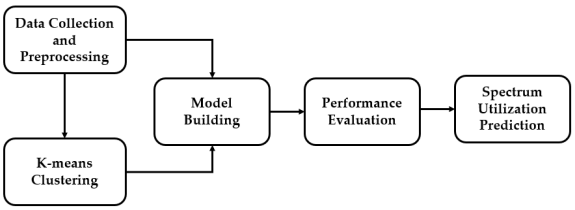

Fig. 2. Overall methodology.

C. K-means Clustering

The K-means algorithm is a partition clustering algorithm. The K variable represents the number of groups for the partition. It works by iteratively assigning data points to one of the K groups based on the provided features. Each data point is assigned to one unique.The Elbow Method is one of the most popular methods to determine this optimal value of K. The optimal number of clusters, K, value is estimated by iteratively observing the inter cluster inertia [13].

1) Clustering Based on Average Spectrum Utilization: The cell load varies constantly due to the differences in human behavior over time and in different locations. Spectrum utilization level clustering means that the average utilization level varies between different base stations. This means that some base stations may have high utilization where others may have low utilization.

Depending on Average spectrum utilization three clusters were obtained. The clustering result and their spatial distribution for the respective three clusters are shown in Fig.2.

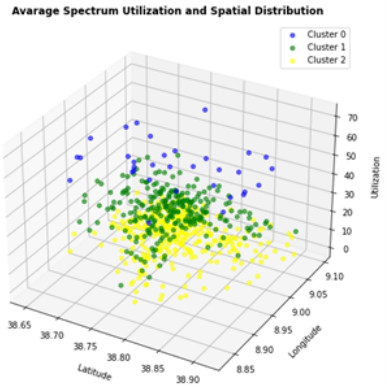

Fig. 3. Clustering Result and their Spatial Distribution.

2) Clustering Based on Temporal Behavior of Spectrum Utilization: Spectrum utilization temporal behavior clustering means that utilization varies during different times and groups. Some base stations may have high utilization during the day, and some may have high utilization during the night period.

Using preprocessed data and the elbow method, the K-means algorithm clusters the data into four clusters based on their temporal behavior. The spectrum usage pattern for the respective four clusters is shown in Fig.4, indicating a presence of diversity and similarity of spectrum usage among them.

Cluster 0 base stations have high spectrum utilization during the day, while cluster 1 base stations have high usage at night.Cluster 2 is heavily used throughout morning and night, while Cluster 3 is heavily used in the morning.

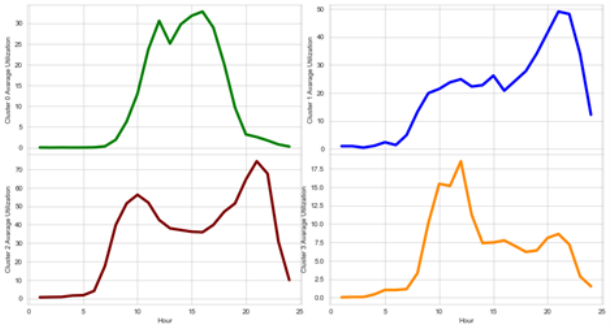

Fig. 4. Daily Spectrum Utilization Pattern of Four Clusters.

The larger the training data set, the better the performance in time series forecasting. But, in the case of the data encountering different behaviors, this may not improve the performance [9]. Hence, the base stations that have similar traffic loads need to be trained together. In this paper, clustering is used to group the base stations with similar spectrum utilization levels and spectrum utilization temporal behavior together. For comparison purposes model is developed using LSTM and CNN algorithms for clustered and non-clustered data.

D. Spectrum Utilization Modeling using LSTM

Long Short Term memory network is an advanced recurrent neural network (RNN) and is capable to learn order dependence in sequence prediction. The LSTM contains three parts, namely Foregate gate, Input gate and Output gate and each part performs a separate function. Complex non-linear feature interactions can be modeled using the LSTM technique [10]. Forget gate chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. Input gate is used to quantify the importance of the new information carried by the input. Output gate the cell passes the updated information from the current timestamp to the next timestamp.

$$f_t = \sigma\left(X_t * U_f + H_{t-1} * W_f\right) \tag{1}$$

$$i_t = \sigma\left(X_t * U_i + H_{t-1} * W_i\right) \tag{2}$$

$$C_t = f_t * C_{t-1} + i_t \tan h\left(x_t * U_c + H_{t-1^*} W_c\right) \tag{3}$$

$$O_t = \sigma\left(X_t * U_o + H_{t-1} * W_o\right) \tag{4}$$

$$H_t = O_t * \tan h\left(C_t\right) \tag{5}$$

E. Spectrum Utilization Modeling using CNN

CNN is a type of deep neural network initially designed for image processing problems, but now it is applied to data that can be represented in a grid-like matrix form. In CNN, time-series and textual data can be represented by a 1D vector and a 2D matrix can be used to represent the pixels in the image data [11].

CNN has an input layer, hidden layers, and an output layer. To achieve the purpose of extracting features there are three main operations to build a CNN model, namely, convolutional layer, pooling layer, and fully connected layer.

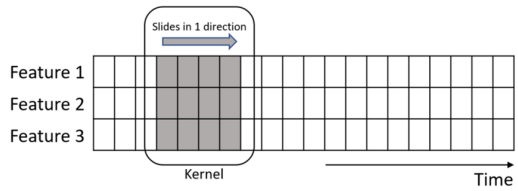

Fig. 5. 1D CNN for Time Series Data.

CNN 1D convolutions to extract information along the time dimension. Convolution can be seen as applying and sliding a filter over the time series. Unlike images, the filters exhibit only one dimension (time) instead of two dimensions.

III. RESULT AND DISCUSSION

Even though the data set was gathered from 639 base stations (1917 cells), five base stations with a continuous missing value were excluded. The data set is divided into 80% of test data, 10% of validation data, and 10% of test data.

A. Hyperparameter tuning

Hyperparameter tuning is finding the best hyperparameters to get the best results from your models. Hyperparameters are set before training a machine learning model. These hyperparameters need to be optimized to adapt a model to a dataset [12].

When building an LSTM model, it is required to consider how many hidden layers the model will include, the number of LSTM cells that should be used in each layer, and what the dropout should be. A grid search algorithm was used to select LSTM model hyperparameters such as a hidden layer, batch size, and epoch. Table 1 shows the hyperparameters used to build the LSTM model.

In addition, The CNN model is tested with various hyperparameters such as kernel size, filter size, hidden layer, optimizer activation function, and Epoch. Table 2 shows the hyper parameters used in CNN model building. Selecting these appropriate combinations of hyperparameters is critical to building a model with better accuracy.

The model will either overfit or underfit the data. The training and the validation loss of the model are plotted to see the possibility of over fitting and underfitting in the model.

TABLE I
HYPERPARAMETERS USED IN LSTM MODEL

Hyperparameters	value
Hidden Layer	3
Hidden layer Neurons	[48,32,32]
Bach Size	128
Dropout	0.2
Optimizer	Adam
Activation	(ReLU)
Epoch	100

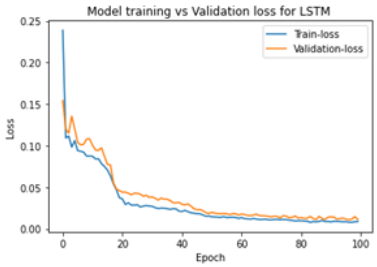

Fig. 6. Plot for Training and Validation Loss of LSTM Model.

B. Model Performance Evaluation Metrics

Two basic prediction metrics are used to evaluate the prediction performance of a constructed model: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

C. Base Station Level Prediction

Base station level prediction is done for one site for 24 hours and the prediction is done using developed models.As shown in the Table IV below, the model developed by CNN algorithm using clustered data can predict up to twenty-four hours of

TABLE II
HYPERPARAMETERS USED IN CNN MODEL

Hyperparameters	value
Number of Filters	64
Kernel Size	[3,3]
Bach Size	64
Hidden Layer	50
Dropout	0.2
Optimizer	Adam
Activation	(ReLU)
Epoch	60

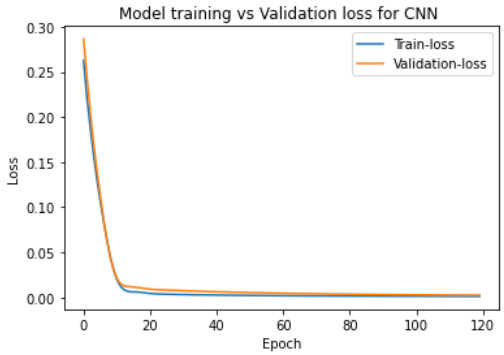

Fig. 7. Plot for Training and Validation Loss of CNN Model.

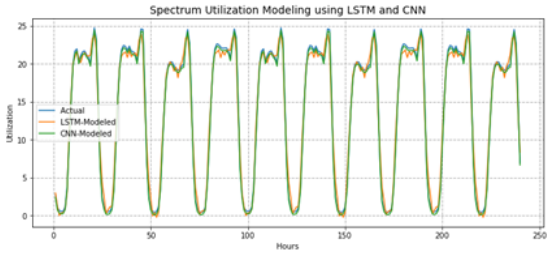

Fig. 8. Actual vs. Modeled Plot for CNN and LSTM.

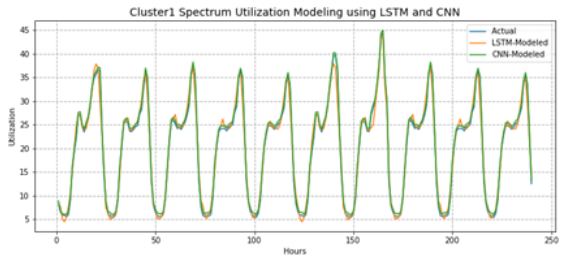

Fig. 9. Cluster data Actual vs. Modeled Plot for CNN and LSTM.

TABLE III
HYPERPARAMETERS USED IN CNN MODEL

	Clustered		Unclustered	
	RMSE	***MAE***	***RMSE***	***MAE***
LSTM	0.8	0.845	1.197	1.057
CNN	0.58	0.26	0.767	0.521

future data traffic with a root mean square error (RMSE) of 1.04.

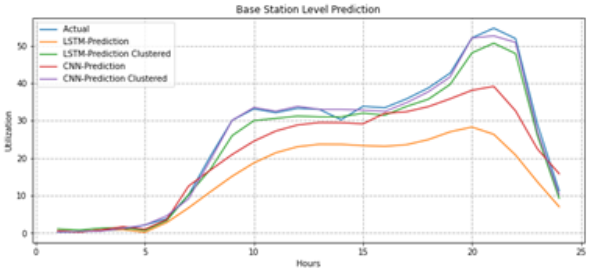

Fig. 10. Base Station Level Prediction Result

TABLE IV
HYPERPARAMETERS USED IN CNN MODEL

	RMSE	MAE
LSTM-Prediction	13.34	10.325
LSTM-Prediction Clustered	2.4492	2.092
CNN-Prediction	7.114	5.058
CNN-Prediction Clustered	1.04	0.76

IV. CONCLUSION

Spectrum utilization data is modeled as time series data during model development, and various strategies for enhancing the model's performance are employed to obtain a better model with the least amount of error. The model is developed using CNN and LSTM algorithms for clustered and non-clustered data. Comparison and evaluation are done using RMSE and MAE. The model developed for the cluster data using the CNN algorithm can model spectrum utilization with an RMSE value of 0.58 and this model after a twenty-four-hour observation of historical data it can predict the next twenty–four–hour base station spectrum utilization with an RMSE value of 1.04.

According to the results, the spectrum is not utilized optimally, so as future work to improve spectrum utilization, apply for dynamic spectrum access and study the impact on spectrum efficiency improvement and its techno-economic effect.

REFERENCES

[1] M. Cave et al., "Review of radio spectrum management," An independent review for Department of Trade and Industry and HM Treasury (www. spectrumreview. radio. gov. uk), 2002.

[2] R. Hafez and G. Chan, "Measures of the spectrum utilization," in VTC'98. 48th IEEE Vehicular Technology Conference. Pathway to Global Wireless Revolution (Cat. No. 98CH36151), IEEE, vol. 1, 1998, pp. 277–281.

[3] P. Dalela, A. Nayak, V. Tyagi, and K. Sridhara, "Analysis of spectrum utilization for existing cellular technologies in context to cognitive radio," in 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011), IEEE, 2011, pp. 585–588.

[4] A. Agarwal, A. S. Sengar, and R. Gangopadhyay, "Spectrum occupancy prediction for realistic traffic scenarios: Time series versus learning-based models," Journal of Communications and Information Networks, vol. 3, no. 2, pp. 35–42, 2018.

[5] M. S. Saitwal, "Dynamic spectrum sharing between mobile network operators in gsm," Ph.D. dissertation, Indian Institute of Technology Hyderabad, 2015.

[6] J. A. Musey and B. Keener, The Spectrum Handbook 2018. 2018.

[7] S. D. Barnes, P. J. Van Vuuren, and B. Maharaj, "Spectrum occupancy investigation: Measurements in south africa," Measurement, vol. 46, no. 9, pp. 3098–3112, 2013.
[8] L. Pedraza, C. Hernandez, and E. Rodriguez, "Modeling of gsm spectrum based on seasonal arima model," in Proceedings of the 6th IEEE Latin-American Conference on Communications, Cartagena, Colombia, 2014, pp. 5–7.
[9] T. A. Rashid, P. Fattah, and D. K. Awla, "Using accuracy measure for improving the training of lstm with metaheuristic algorithms," Procedia Computer Science, vol. 140, pp. 324–333, 2018.
[10] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long shortterm memory (lstm) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132 306, 2020.
[11] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," IEEE transactions on visualization and computer graphics, vol. 23, no. 1, pp. 91–100, 2016.
[12] J. Jordan, "Hyperparameter tuning for machine learning models," Retrieved from: Jeremy Jordan: https://www. jeremyjordan. me/hyperparameter-tuning, 2017.
[13] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol. 336, 2018, p. 012 017.