MACHINE LEARNING MODELS FOR CHANNEL STATUS CLASSIFICATION IN M-MIMO SYSTEMS USING LIMITED CSI FEEDBACK

Ben Earle
Ala'a Al-Habashna
Gabriel Wainer

Carleton University
1125 Colonel By Dr
Ottawa, ON K1S 5B6, CANADA
BenEarle@cmail.carleton.com,
{alaaalhabashna, gwainer}@sce.carleton.ca

Xingliang Li
Guoqiang Xue

Ericsson
349 Terry Fox Drive,
Ottawa, ON K2K 2V6, CANADA
{xingliang.li, guoqiang.xue}@ericsson.com

ABSTRACT

In next generation networks, knowing if a channel has a Line of Sight (LOS) path between the transmitter and the receiver is becoming increasingly important. For example, researchers have optimized channel estimation and wireless localization algorithms for both LOS and Non-LOS (NLOS) scenarios. Knowing the LOS status of a channel will allow system performance enhancement by employing the best algorithm available. This study explores the use of various machine learning classifiers to identify the LOS status of simulated massive-MIMO channels. The classifiers make their predictions based on limited Channel State Information (CSI) feedback received at the base station. This study identifies and properly manages the class imbalance problem present in LOS/NLOS identification. Promising results are achieved and demonstrated using a synthetic benchmark.

Keywords: Machine Learning, Neural Networks, Wireless Channel, LOS identification.

1 INTRODUCTION

Wireless communications technology has been witnessing an exponential growth in the requirements with each generation (ITU-R 2015). Massive Multi Input Multi Output (m-MIMO) antenna arrays are becoming a necessary technology to meet the increasing user count, data rate, and reliability expectations in next generation networks (Björnson, Hoydis, and Sanguinetti 2017). These systems accomplish this through improved beamforming, which increases signal strength and improves spectral efficiency using spatial multiplexing. Channel Estimation (CE) is a major challenge for m-MIMO systems, since the channel matrix size is multiplicative based on antenna count. Knowing if a channel has a Line of Sight (LOS) path between the transmitter and the receiver allows the devices to optimize their transmissions. CE algorithms are an example of an algorithm which can perform better based on the LOS status of the channel. Furthermore, User Equipment (UE) localization is a topic of rising interest due to novel use cases in 5G systems (Ericsson 2020). Many localization algorithms also benefit from knowing if the channel has a LOS path or is entirely Non-LOS (NLOS), ultimately making LOS/NLOS identification an asset to optimize wireless systems.

This paper explores the use of Machine Learning (ML) classification algorithms to identify the LOS status of a given channel. The data is generated using a MATLAB simulator with a combination of custom and

ANNSIM'23, May 23-26, 2023, Mohawk College, ON, CANADA;

5G Toolbox functions. The simulator uses the 3GPP Clustered Delay Line (CDL) channels (3rd Generation Partnership Project 2020), which are statistical models which assume the channel is made up of several clusters of reflectors. Each channel will have several clusters, each cluster will have several paths with similar path parameters. The channels are implemented with 5G Toolbox and the cluster parameters are created using a custom generation script. Channel State Information (CSI) metrics are calculated at the UE based on simulated reference signals. Limited CSI are fed-back to the Base Station (BS) and used by the ML models to predict the LOS/NLOS status of the channel. A data frame that contains simulation CSI and the ground truth value for LOS/NLOS status is used as the supervised learning data set for training and testing ML models. The following ML algorithms are used in this study: K-Nearest Neighbours (KNN), Random Forest (RF), several different Neural Network (NN) architectures, and a meta-learning solution.

Similar studies have been conducted in this space, however they all focus only on classification accuracy or error as their performance metric. In practice there are many scenarios (such as, urban or indoor) where most of the wireless channels are NLOS. The imbalance between LOS and NLOS cases invalidates the use of accuracy as a sole performance metric. Instead of accuracy, this study uses the Area Under the Curve (AUC) of a Precision-Recall (Pr-Re) plot. Pr-Re plots focus on the minority class and provide two metrics which show the complete performance of the classifier. The AUC of a Pr-Re curve is the preferred metric for reporting on generic classifiers in the presence of class imbalance. The performance of several ML classifiers is shown on three different simulated scenarios. Additionally, their performance is analyzed on a synthetic model which emulates a real world scenario. Their performance in this synthetic benchmark exceeds that of a classifier that always reports the dominant class (NLOS), despite having lower accuracy.

The remainder of this paper is organized as follows. First, Section 2 provides the background on the usefulness of LOS identification, the channel models used, related work, and a high level description of the ML algorithms used in the study. Section 3 describes the simulator structure and data sets used. Section 4 discusses the implementation details of each of the ML models. Section 5 presents the results, which include Pr-Re curves for our best performing models. Finally, Section 6 provides the conclusion and future work.

2 BACKGROUND

Transmitted wireless signals are affected by the physical environment when propagating to the receiver. The wireless channel will distort the signal, altering the gain and phase of the wireless waves. Furthermore, the waves can reflect and refract off surfaces, causing multiple copies of the signal to arrive at the receiver at separate times, different powers, and out of phase. These phenomena are modeled by the CDL channel model defined in (3rd Generation Partnership Project 2020); which is a multi-path clustering based statistical model. A cluster is made up of several rays (individual paths between transmitter and receiver) which are all traveling in the same general direction. A given reflector will have several rays that represent the same cluster. The cluster is defined by a delay, average gain and angle. The delay and gain are relative to the first cluster to arrive and the highest cluster gain. Each ray within the cluster will have essentially the same delay, and a slightly different gain and angle. The CDL model allows custom channels to be created, where the user may select the channel parameters. Channel parameters are broken into two categories: Large Scale Parameters (LSPs), and Small Scale Parameters (SSPs). LSPs are determined by the physical environment, such as distance between BS and UE, the LOS status, the LOS angle, and the path loss. LSPs are shadow fading, K-factor, delay spread, and azimuth and zenith angle spread for arrival and departure angles. SSPs are random values calculated using the LSPs. They are the individual cluster delays, powers, azimuth and zenith angles of arrival and departure, and the cross power ratios. The SSPs are then used when calculating the channel matrix as a function of time.

Channels are often modeled as a complex matrix, with each element representing the delay and gain that a given transmitter will experience at the receiver (Rappaport 2001). Channel sounding is the process where

a known signal is transmitted so that the receiver can calculate the channel's effect on the signal. The BS controls the communication protocols and transmission parameters used (encoding scheme, data rate, etc.), where the optimal protocols and parameters are dependent on the physical channel. When sounding the channel between the BS and UE, the UE must estimate the channel and return the results to the BS. Any resources spent sending control data, such as the channel estimate, are resources lost that could have been spent sending data. Therefore, to minimize this overhead, the UE calculates Channel State Information (CSI) metrics to send to the BS, instead of sending a full channel estimate. In practice the BS must then select its transmission protocols based on this fed-back CSI; for that reason, CSI is the chosen input for the machine learning models in this study. The models will use the CSI data to predict if the channel has a LOS cluster, aiding in the BS algorithm and protocol selection. Specifically, the models are limited to the following CSI: Received Signal Strength Indicator (RSSI), Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Channel Quality Indicator (CQI), Precoding Matrix Indicator (PMI), and Rank Indicator (RI). RSSI, RSRP, and RSRQ are all reference powers in Db or DBm. The CQI is an index into a predetermined list of signal to noise ratios. The PMI is an index into a list of matrices that will tell the BS what gain and phase shift to give to each antenna to maximize the received signal. Finally, the RI indicates the number of orthogonal paths to the receiver.

Knowing if the channel has a LOS path allows the BS to select the best algorithms in different parts of the system, such as CE or localization. For instance, the geometrical channel model proposed (Liberti and Rappaport 1996) is only valid for LOS channels, as is the delay domain channel model from (Kyro, Kolmonen, and Vainikainen 2012) for mm-wave channels, and the CE strategy from (Ji, Fan, and Pedersen 2017) for indoor spherical wave channels. Similarly, if a system is known to be NLOS then it could use the advanced CE method in (Milenkovic, Panic, Denic, and Radenkovic 2017) for mm-wave, or the optimized CE algorithm presented in (Zhang, Gong, and Xu 2014) for optical wireless channel estimation. Similar optimizations can be made to wireless localization problems. (Adebomehin and Walker 2016) proposes an algorithm for LOS localization, and (Fan, Chu, Wang, and Lu 2020), (Yang, Li, and Ye 2016), (Cheng et al. 2017) and (Wang, Cheng, and Hu 2015) all propose wireless sensor localization algorithms for the more-probable NLOS scenario. LOS/NLOS identification is prerequisite to being able to use any of the aforementioned algorithms, and thus is an important problem to solve.

This study uses exclusively supervised learning, meaning that the ML algorithms were trained and tested on labeled data sets. The models used were KNN, RF, NN, and a meta-learning solution. KNN is a simple algorithm used in this work as a baseline to show how well other solutions are performing. When classifying a data point, KNN uses the training data features to find the k closest training labels and averages their classes to make a prediction (Hastie, Tibshirani, and Friedman 2009). In these models k is an integer and the optimal value is determined while training. A Decision Tree (DT) is essentially a flow chart where each node checks the value(s) of a specific feature(s) and then progresses in a different direction depending on the result. Each layer will have some kind of branching conditions leading to the next layer. In the final layer of the tree, a classification is assigned. RF is a set of many un-correlated DTs, when classifying a data point the RF will average the result of all the DTs. A NN is a network model which consists of neurons (Burkov 2019). Each neuron does a linear combination of its inputs along with training weights, then decides based on the activation function what it should output. When designing a NN, one can vary the width (number of nodes per layer) and the depth (number of layers in the network). During training, the network will optimize the weights in each neuron. In a classification problem, like the one presented in this paper, the final layer of the network will output values from 0-1, where each value indicates the probability that the input data belongs to certain class in the classification problem.

ML and statistical modeling has been used for LOS identification in several other studies. (Zhang, Salmi, and Lohan 2013) proposed LOS/NLOS identification for localization purposes using kurtosis of the amplitude of the channel impulse response. (Wang et al. 2019) used Convolutional NN (CNN) on the channel impulse response to identify LOS status with a very high accuracy for indoor ultra-wide band scenarios. (Zeng et al.

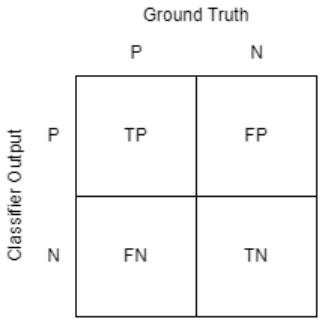

Figure 1: Confusion matrix structure.

2018) also used a CNN, the features in their model are the m-MIMO antenna powers during up-link channel sounding. Their reported accuracy is 97.5%. (Huang et al. 2020) used support vector machine, random forest, and NN to predict LOS/NLOs conditions from the channel impulse response. They also report an accuracy of 99% in their test cases. (Carpi et al. 2019) used the values from several consecutive channel estimates with NN classifiers and other statistical estimators to predict the LOS status of the channel. They designed their system for IEEE 802.11 links and reported an accuracy of 85%-90% for their models. These studies predicted LOS status to assist CE or UE localization, using different statistical models and features. One thing that they all had in common was that they used accuracy or raw error rate as their only reported metric. These studies did not report other metrics, or provide a confusion matrix for their classifier. This is an important issue because a majority of urban and indoor channels (which are used in these studies) will have a much higher NLOS probability due to their environments. This creates a class imbalance problem for LOS/NLOS identification which is not addressed in previous works.

Data which has a disproportionately large number of one class (normally the negative class in a binary classification problem) within the population are said to have class imbalance. Typically, in these scenarios the purpose of the model is to identify the rare class, so it is important that it is not overlooked (for example, finding fraudulent charges on a credit card, or diseased patients in a medical scan). Class imbalance, when not dealt with properly, can cause the statistical model to favor the dominant class and skew accuracy metrics. For instance, if 99% of the channels evaluated are NLOS, and we classify all channels as NLOS, we will achieve 99% prediction accuracy without actually gaining any knowledge about the channel. In studies that deal with class imbalance, it is better to report the raw results of the classifier in a confusion matrix, shown in figure 1, since it demonstrates the complete performance of the classifier (Nandi and Ahmed 2020).

The horizontal axis shows the true class and the vertical axis the predicted class. The True Positives (TP) are instances where the classifier correctly assigns the positive class to the data point. False Positives (FP) are instances where the classifier predicted positive but the actual value is negative. True Negatives (TN) and False Negatives (FN) are defined similarly, only the class is reversed. Accuracy is defined as follows,

$$Accuracy = \frac{TP + TN}{N}, \tag{1}$$

where N is the total number of samples in the population, making accuracy the fraction of the population that was correctly classified. This becomes a problem when only a small percentage (e.g., 1%) of our data belongs to the positive class. In this scenario, the classifier could predict that every sample is negative, the TP's would be zero, and the classifier would still achieve an accuracy of 99%. Alternatively, more meaningful metrics for this type of problem are Precision and Recall, defined as,

Table 1: Simulation Parameters and Scenarios.

(a) Input Parameters

Input	Value
Number of Channels	5000
Scenario	UMa
Max Delay Spread	1000ns
Carrier Frequency	2GHz
Transmitter Antennas	32
Receiver Antennas	4

(b) Scenarios

Scenario	Cell Radius	LOS Channels	LOS %
S1	5,000m	47	0.94%
S2	1,000m	238	4.76%
S3	200m	1605	32.1%

$$Precision = \frac{TP}{TP+FP}, \tag{2}$$

$$Recall = \frac{TP}{TP+FN}. \tag{3}$$

Precision is the fraction of positive data points that are correctly classified. Recall is the fraction of positive classifications that were correct. These two values are inversely related and dependent on the classification threshold. Binary classifiers output a value between 0 and 1, the higher the value the more likely the classifier believes the inputs to be part of the positive class. A classifier that is very strict will have a higher classification threshold. This means only classifying data as positive when it is very certain, which will achieve a very high precision; since there will be several TPs and very few FPs. However this high precision classifier will miss many of the positive classes, increasing the FN count and driving down the Recall. Likewise, a more lenient classifier, with a lower classification threshold, will favor the positive class. The lenient classifier will have a high TP, low FN, and high FP rate; thus driving up the recall and down the precision. Plotting the Precision against the Recall at each classification threshold between zero and one shows the classifier performance at various operating points. The ideal operating point depends on the specific application (is a FP or FN more costly?). Without knowing the exact cost of an incorrect classification the Pr-Re AUC acts as an ideal metric for generic classifier performance when a class imbalance is present.

3 SIMULATOR DESIGN AND SCENARIO PARAMETERS

The channel simulator is an improved version of the one used in (Earle et al. 2021). It is designed to simulate as many 3GPP CDL channels as described in (3rd Generation Partnership Project 2020) with realistic and custom channel parameters, and log the conditions of the simulation and CSI fed back to the BS. Currently, only Urban Macro channels have been considered. However, the simulator supports Urban Micro channels and work is underway to add indoor channels as well. The primary parameter that was adjusted in this study was the size of the cell, as this directly impacted the LOS probability for a given channel. The remainder of the input parameters, along with their values, are shown in table 1 (a). Three data sets were simulated called S1, S2, and S3 using the above conditions and cell size of 5,000m, 1,000m, and 200m, respectively. The resulting data distribution of each of these data sets is shown in table 1 (b). S1 is a realistic size for an UMa cell, S2 is small, and S3 is unrealistically small. Running the simulator for 5000 channels took approximately 8 hours, so the channel size was decreased to artificially increase the number of LOS channels in the data set to improve the models ability to identify the LOS case.

The simulator architecture is shown in figure 2. The channel parameter generation is defined in (3rd Generation Partnership Project 2020). All inputs shown are adjustable, however, they were set to the shown values in this study. The parameter generation follows the definition in (3rd Generation Partnership Project

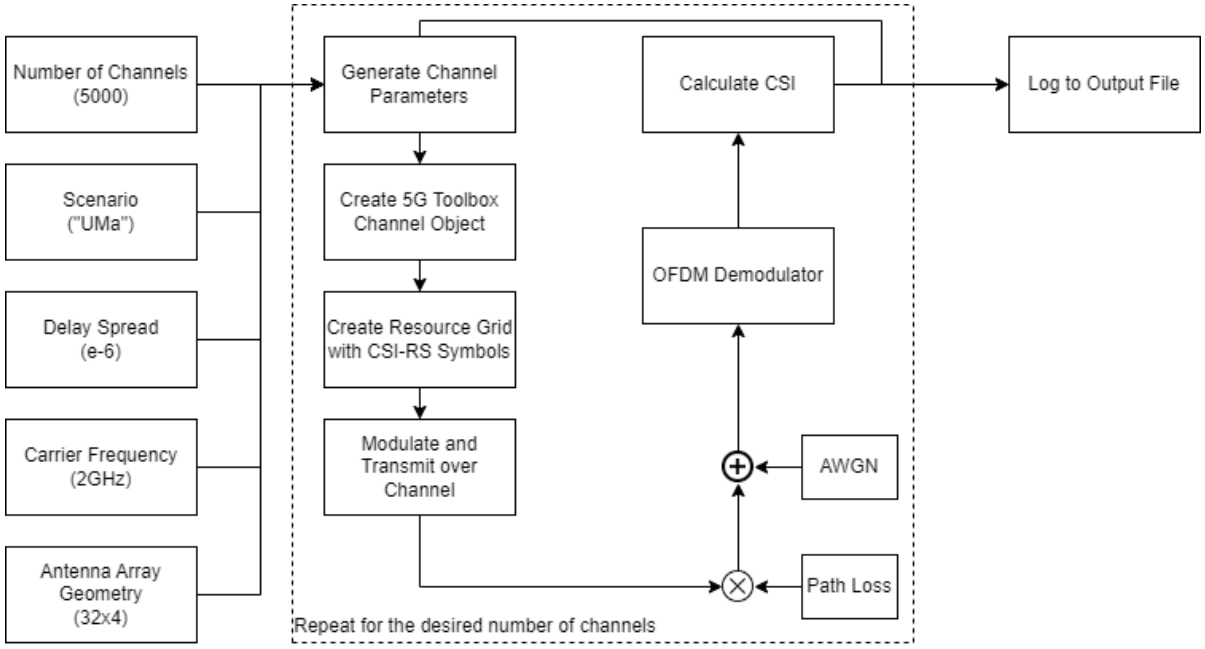

Figure 2: Simulation Architecture.

2020); the process was implemented and explained in (Riviello, Di Stasio, and Tuninato 2022), along with their MATLAB code being available on gitlab (Riviello 2022). Their parameter generation script was used as a starting point and significantly updated to meet the parameter generation needs of this project. The parameters were used to create custom CDL channels using the 5G Toolbox, which handled the resource grid creation, modulation, and application of the channel effects. Afterwards, path loss is applied along with additive white Gaussian noise. Then, a combination of custom and 5G Toolbox functions are used to calculate the CSI based on the received CSI-RS signal. The CSI and the channel parameters are logged to be used by the ML models.

4 MACHINE LEARNING MODELS

This section will include the data pre-processing, the hyper-parameters used for the models in this study, and the methods used to manage class imbalance. Many combinations of pre-processing and class imbalance management were tested to find the optimal results. The results from all combinations are presented in section 5, along with Pr-Re curves and a theoretical example use-case for the best performing models.

Data pre-processing is important for efficiently training models that are sensitive to the absolute value of their inputs, such as NN. If some features have larger values, like RSSI, then their updates will be larger and they will have artificially higher impact on the output until the model learns to decrease their weight accordingly. This problem is prevented through scaling or standardizing the data. In this work we tested two scalers, the first was l2 normalization (Scikit-Learn b), which was abbreviated to NRM in resulting Pr-Re figures. The second is the Min Max Scaler (MMS), which subtracts the minimum value for the feature from each entry then divides by the difference between the minimum and maximum values (Scikit-Learn a). This results in a scaled value between zero and one. Scalers change the magnitude but preserve the feature data's original distribution; whereas standardizers will change the values and distribution. We tested one standardizer, which removed the mean and scaled the values to have unit variance, abbreviated to SS. The NRM, MMS, and SS data pre-processors were tested on all the models without any data balancing on the S3 data set with 50 epochs. This was chosen since S3 has an artificially high positive count. The best

Table 2: NN Architectures.

N1

Layer	Neurons	Activation
Input	33	Relu
Hidden 1	32	Relu
Hidden 2	8	Relu
Output	1	Sigmoid

N2

Layer	Neurons	Activation
Input	33	Relu
Hidden 1	128	Relu
Hidden 2	32	Relu
Output	1	Sigmoid

N3

Layer	Neuron	Activation
Input	33	Relu
Hidden 1-4	128	Relu
Hidden 5-8	64	Relu
Hidden 9-12	32	Relu
Hidden 13-16	16	Relu
Hidden 17	8	Relu
Output	1	Sigmoid

N4

Layer	Neuron	Activation
Input	33	Relu
Hidden 1-4	128	Relu
Hidden 5-8	64	Relu
Hidden 9-12	32	Relu
Hidden 13-16	16	Relu
Hidden 17	8	Relu
Output	1	Sigmoid

N5

Layer	Neuron	Activation
Input	5	Relu
Hidden 1	16	Relu
Hidden 2	8	Relu
Output	1	Sigmoid

performing pre-processor was MMS as it had the highest average and maximum Pr-Re AUC so it was used for all future analysis.

Machine learning models have difficulty in identifying the non-dominant class in the presence of class imbalance. A common way to improve the model's ability to identify the minority class is to balance the training data. It is important to note that the training and validation data may be balanced but the final hold-out test must not be, to reflect the actual performance on the population. Three methods of training data balancing were tested in this study: Over-sampling, Under-sampling, and Synthetic Minority Oversampling Technique (SMOTE). The goal of data balancing is to have an equivalent amount of LOS and NLOS channels in the training data set so the optimization algorithm values them evenly. Over-sampling is when data from the minority class is increased (either by duplicating some samples or creating synthetic samples). Under-sampling is when fewer data points from the majority class are considered. SMOTE is an algorithm that generates fake data points for the minority class that match its feature distribution. These algorithms were tested on all three scenarios since their level of imbalance is different. Overall, not balancing the training data performed the best for all three data sets. This was not the expected outcome. Although it is expected that this happened because all models used the Pr-Re AUC as their performance metric during training. The models trained on balanced data over predicted the LOS case causing the Pr-Re curves to be worse than those trained with the real data distribution when testing on the hold-out data.

There were seven different models tested and tuned for this study. The first two are RF and KNN models, the remaining five are NN of varying architectures. The data was partitioned to be 70% for training and validation, and %30 as a hold-out test. The RF used the validation data to optimize the hyper-parameters, typically the optimal structure had 100 DTs voting on the classification. Likewise, the validation set was used to tune the KNN model, and typically the model used the 5-7 nearest neighbours. Note that the Pr-RE curves for the KNN model are not smooth, since the number of neighbours used is low. The NNs are numbered N1-N5 and their architectures are shown in table 2. The networks each represent a specific architecture

style: N1 is shallow and narrow, N2 is deep and narrow, N3 is shallow and wide, and N4 is both Deep and Wide. N5 is a meta-learning network; instead of using the CSI values like the other models, its inputs are the outputs from the RF model and N1-N4. It does not use the KNN output because its performance was relatively poor and it only made N5's decision making worse overall.

5 MODEL VALIDATION AND SYNTHETIC BENCHMARK

The first series of tests conducted were to determine the best pre-processing methodology. The S3 data set was used since it had a reasonable number of LOS and NLOS samples. The number of epochs was held constant at 50 while varying the pre-processing algorithm. The results of this test are shown in table 3 (a). The method, average AUC, best model and the best model's AUC are all presented. Overall, pre-processing algorithm did not seem to have a major impact on this data. MMS was used for all future testing since it tied others for average AUC and the best AUC recorded was using MMS. Testing was also done to determine the best data-balancing technique. These tests were conducted for all three scenarios, since each scenario has different degrees of class imbalance. The results are shown in table 3 (b). Interestingly, balancing the data did not seem to help in any of the scenarios. All three found that training on the unbalanced data achieved better Pr-Re AUCs on the hold-out test. Additionally, these results show that the scenarios with more class imbalance favor the meta-learning models (N5 and RF). The final test conducted was to see how the number of epochs affected the models results. The results are shown in table 3 (c). These tests were done with MMS and no training data balancing for all three scenarios. Overall, these results show an improvement in average and best AUC as the number of epochs approach 50, then they start to perform worse, since the algorithms over-fit with too many training epochs. In this batch of tests, the meta-learners continued to outperform the NN and KNN models. The best performing model's Pr-Re curve for each scenario is shown in figure **??**.

In a real world application, the system would have three choices based on the LOS/NLOS classifier output. The system may use an LOS or NLOS specific algorithm, or it could use a generic algorithm. Therefore, only the classifications above a certain confidence level need to be considered. Applying this information to the problem, the following equation is used when assigning classes to the classifiers output,

$$\begin{aligned} &1 : prediction > 1 - threshold \\ &0 : prediction < threshold \\ &? : otherwise \end{aligned} \tag{4}$$

In a hypothetical example, consider a system that has a generic localization algorithm which has a 10m accuracy. The system may use a LOS specific algorithm with a 1m accuracy when properly applied to LOS channels and 50m accuracy, when incorrectly applied to NLOS channels. The system also has an NLOS specific algorithm which has a 5m accuracy when applied to NLOS channels and 50m accuracy when applied to LOS channels. Several thresholds were tested with the optimal configuration described previously. The results of the best performing algorithm for each scenario are presented in table 4. The performance is the average estimation accuracy for the synthesized scenario. The TP, FP, FN, and TN are the values in the confusion matrix for the data points that were classified with the confidence threshold. The Acc, Pr, and Re columns are the Accuracy, Precision, and Recall, respectively. The final column, class, is the ratio of data points that were given a class instead of left to the unknown class. There are two baselines that the ML models need to outperform. The first is "All ?", which is the result when treating all the channels as the unknown LOS status. The second baseline is specific to the scenario; it is what happens if all channels are given the NLOS class. The second baseline is more difficult to beat in S1 since it has the largest class imbalance. In S1, the baseline that must be beat is 5.42m, the best performing model was RF with a 5.32m estimation accuracy, and the optimal result was 5.0m. In S2, the baseline was 7.1m, the best

Table 3: Hyper Parameter Tuning

(a) Pre-Processing Evaluation.

Method	Average AUC	Best Model	Best AUC
NRM	0.81	N2, N4	0.84
SS	0.79	N2, N4	0.82
MMS	0.81	RF	0.86
None	0.81	RF, N3, N4	0.84

(b) Training Data Balancing Evaluation.

Scenario	Method	Average AUC	Best Model	Best AUC
S1	ROS	0.19	N5	0.44
S1	RUS	0.06	N5	0.23
S1	SMOTE	0.13	N5	0.28
S1	None	0.33	N5, RF	0.50
S2	ROS	0.37	RF, N5	0.6
S2	RUS	0.28	N5	0.38
S2	SMOTE	0.38	N5	0.47
S2	None	0.45	RF, N5	0.61
S3	ROS	0.79	N2, N4	0.83
S3	RUS	0.78	N2, N4	0.81
S3	SMOTE	0.73	RF	0.84
S3	None	0.81	RF, N3, N4	0.86

(c) Epoch Evaluation.

Scenario	Epochs	Average AUC	Best Model	Best AUC
S1	10	0.30	N5	0.52
S1	25	0.27	N5	0.45
S1	50	0.24	RF	0.44
S1	150	0.21	RF	0.42
S1	250	0.25	RF	0.42
S1	500	0.31	N5	0.44
S2	10	0.38	RF, N5	0.51
S2	25	0.38	N5	0.51
S2	50	0.39	RF, N5	0.5
S2	150	0.37	RF	0.51
S2	250	0.38	N5	0.52
S2	500	0.37	N5, RF	0.52
S3	10	0.81	RF, N5	0.85
S3	25	0.81	RF, N5	0.85
S3	50	0.80	RF, N5	0.84
S3	150	0.79	RF	0.85
S3	250	0.78	RF	0.85
S3	500	0.77	RF	0.85

performing model was RF at 6.0m, and the perfect estimator was 4.8m. Finally, S3's baseline was the "All" model at 10m, the best model was N1 which achieved 7.7m, and the optimal classifier gets an accuracy of 3.7m. Ultimately, the ML models improved the systems performance when compared to the baseline in the synthetic benchmark scenario. The improvement relative to the baseline was about 2% for S1, 24% for S2, and 29% for S3.

6 CONCLUSION AND FUTURE WORK

This paper explores the use of various ML models to estimate the LOS status of a channel based on CSI feedback. The channels used were custom 3GPP CDL channels. An RF, KNN, and four NNs used RSSI, RSRP, RSRQ, CQI, PMI, and RI to predict the probability a channel had a LOS path. Additionally, a fifth NN was used as a meta-learner to combine the results of the RF and NNs. The study focused on the AUC of the Pr-Re plot to accurately report the classification performance in presence of class imbalance. Finally, a theoretical example was used to show the developed model's potential. The results were promising for a preliminary study and the remainder of this section will explore next steps to improve this research.

In future iterations of this work, more ML models should be tested on larger data sets. The results for S1 and S2 were not as good as they could have been due to the lack of samples of LOS channels to learn from. If larger simulated data sets were used, there would be more positive classes and the results could improve. Additionally, there are many other types of ML classification models which is worth testing. Some examples are Support Vector Machines, Naive Bayes, Boosting, and other NN architectures. Primarily, future work

Table 4: Algorithm Performance in the Synthesized Scenario.

Scenario	Model	Threshold	Performance	TP	FP	FN	TN	Acc	Pr	Re	Class
All	All ?	0	10	0	0	0	0	0	0	0	0
S1	All NLOS	0	5.42	0	0	14	1486	0.991	0	0	1
S1	RF	0.05	5.32	0	0	6	1452	0.996	0	0	0.97
S2	All NLOS	0	7.1	0	0	70	1430	0.95	0	0	1
S2	RF	0.2	6.0	3	0	24	1373	0.98	1	0.11	0.93
S3	All NLOS	0	19.43	0	482	0	1018	0.68	0	0	1
S3	N1	0.1	7.7	176	13	17	600	0.96	0.93	0.91	0.54

should focus on meta-learning strategies, as the two tested in this preliminary research lead to the best results. It would also be beneficial to implement one of the LOS/NLOS identification algorithms discussed in the background to compare with our CSI-based solution for additional perspective. Furthermore, this study can be extended by finding a real world example for the performance gain/cost for TP, FP, TN, and FN classifications. Finally, in our results we found that a substantial amount of the LOS classes had an NLOS path with more power than their LOS path. So a future classifier may benefit from identifying between these two scenarios as well.

CDL channel models are useful for designing and testing a systems performance under generic channel conditions. Future iterations of this work could be extended by including UMi and indoor environments, since a majority of UE localization problems use smaller cells. Improvements to the simulator to add support for these models is already underway. Additionally, a LOS/NLOS identification algorithm would actually benefit from learning the characteristics of the cell they are deployed in. Hypothetically, a cell with a park surrounded by tall buildings will have a much higher LOS probability if a user's signal is coming from the park instead of the buildings. Furthermore, by using online-learning, an ML model could infer direction using the PMI, and improve the prediction results over time. This system would first be trained on the statistical channel model, then have a simulated deployment using a geometric channel model. The hypothesis is that the online-learning model should adapt to its environment and outperform the starting model fairly quickly.

REFERENCES

The 3rd Generation Partnership Project 2020. "Study on channel model for frequencies from 0.5 to 100 GHz TR 38.901".

Adebomehin, A. A., and S. D. Walker. 2016. "Enhanced Ultrawideband methods for 5G LOS sufficient positioning and mitigation". In *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–4.

Björnson, E., J. Hoydis, and L. Sanguinetti. 2017. *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. Foundations and Trends in Signal Processing.

Burkov, A. 2019. *The Hundred-Page Machine Learning Book*.

Carpi, F., L. Davoli, M. Martalò, A. Cilfone, Y. Yu, Y. Wang, and G. Ferrari. 2019. "RSSI-based Methods for LOS/NLOS Channel Identification in Indoor Scenarios". In *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 171–175.

Cheng, L., Q. Qi, X. Wu, Y. Shao, and Y. Wang. 2017. "A NLOS selection based localization method for wireless sensor network". In *2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 340–343.

Earle, B., A. Al-Habashna, G. Wainer, X. Li, and G. Xue. 2021. "Prediction of 5G New Radio Wireless Channel Path Gains and Delays Using Machine Learning and CSI Feedback". In *2021 Annual Modeling and Simulation Conference (ANNSIM)*, pp. 1–11.

Ericsson 2020, Dec. "5G positioning: What you need to know". https://www.ericsson.com/en/blog/2020/12/5g-positioning--what-you-need-to-know. Accessed Mar. 01, 2023.

Fan, Z., H. Chu, F. Wang, and J. Lu. 2020. "A New Non-Line-of-Sight Localization Algorithm for Wireless Sensor Network". In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pp. 858–862.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Huang, C., A. F. Molisch, R. He, R. Wang, P. Tang, B. Ai, and Z. Zhong. 2020. "Machine Learning-Enabled LOS/NLOS Identification for MIMO Systems in Dynamic Environments". *IEEE Transactions on Wireless Communications* vol. 19 (6), pp. 3643–3657.

ITU-R 2015, Sep. "IMT Vision 2020". https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf. Accessed Mar. 28, 2023.

Ji, Y., W. Fan, and G. F. Pedersen. 2017. "Channel estimation using spherical-wave model for indoor LoS and obstructed LoS scenarios". In *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pp. 2459–2462.

Kyro, M., V.-M. Kolmonen, and P. Vainikainen. 2012. "Experimental Propagation Channel Characterization of mm-Wave Radio Links in Urban Scenarios". *IEEE Antennas and Wireless Propagation Letters* vol. 11, pp. 865–868.

Liberti, J., and T. Rappaport. 1996. "A geometrically based model for line-of-sight multipath radio channels". In *Proceedings of Vehicular Technology Conference - VTC*, Volume 2, pp. 844–848 vol.2.

Milenkovic, V., S. Panic, D. Denic, and D. Radenkovic. 2017. "Novel method for 5G systems NLOS channels parameter estimation". *International Journal of Antennas and Propagation* vol. 2017, pp. 1–5.

Nandi, A. K., and H. Ahmed. 2020. *Condition monitoring with vibration signals compressive sampling and learning algorithms for rotating machines*. Wiley-IEEE Press.

Rappaport, T. S. 2001. *Wireless communications: principles and practice*. Pearson.

Riviello, D., F. Di Stasio, and R. Tuninato. 2022, 01. "Performance Analysis of Multi-User MIMO Schemes under Realistic 3GPP 3-D Channel Model for 5G mmWave Cellular Networks". *Electronics* vol. 11.

Riviello, Daniel Gaetano 2022. "3GPP channel model TR 38901". https://gitlab.com/daniel.riviello/3gpp-channel-model-tr-38901. Accessed Mar. 01, 2023.

Scikit-Learn. "Sklearn.preprocessing.MinMaxScaler". https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Accessed Mar. 01, 2023.

Scikit-Learn. "Sklearn.preprocessing.normalize". https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html. Accessed Mar. 01, 2023.

Wang, F., Z. Xu, R. Zhi, J. Chen, and P. Zhang. 2019. "LOS/NLOS Channel Identification Technology Based on CNN". In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 200–203.

Wang, Y., L. Cheng, and N. Hu. 2015. "Bayes sequential test based NLOS localization method for wireless sensor network". In *The 27th Chinese Control and Decision Conference (2015 CCDC)*, pp. 5230–5234.

Yang, Y., B. Li, and B. Ye. 2016. "Wireless Sensor Network Localization Based on PSO Algorithm in NLOS Environment". In *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Volume 01, pp. 292–295.

Zeng, T., Y. Chang, Q. Zhang, M. Hu, and J. Li. 2018. "CNN-Based LOS/NLOS Identification in 3-D Massive MIMO Systems". *IEEE Communications Letters* vol. 22 (12), pp. 2491–2494.

Zhang, J., J. Salmi, and E.-S. Lohan. 2013. "Analysis of Kurtosis-Based LOS/NLOS Identification Using Indoor MIMO Channel Measurement". *IEEE Transactions on Vehicular Technology* vol. 62 (6), pp. 2871–2874.

Zhang, X., C. Gong, and Z. Xu. 2014. "Estimation of NLOS optical wireless communication channels with laser transmitters". In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pp. 268–272.

AUTHOR BIOGRAPHIES

BEN EARLEis a PhD student in Computer Engineering at Carleton University under the supervision of Dr. Gabriel Wainer and Dr. Ala'a Al-Habashna. His research interests include wireless communication, embedded systems, modeling and simulation, and machine learning. His email address is benearle@cmail.carleton.ca.

ALA'A AL-HABASHNA received his Master of Engineering degree from Memorial University of Newfoundland in 2010, and his PhD degree from Carleton University in 2018, both in Electrical and Computer Engineering. Currently, Dr. Al-Habashna is an Adjunct Research Professor at Carleton University and a senior researcher at Statistics Canada, Ottawa, Canada. His current research interests include 5G wireless networks, IoT applications, multimedia communication over wireless networks, discrete-event modeling and simulation, signal detection and classification, cognitive radio systems, and applied machine learning and computer vision. His email is alaaalhabashna@sce.carleton.ca.

GABRIEL WAINER is a Professor in the Department of Systems and Computer Engineering, Carleton University (Ottawa, ON, Canada). His current research interests are related with modeling methodologies and tools, parallel/distributed simulation, and real-time systems. He is a Fellow of SCS. His e-mail is gwainer@sce.carleton.ca. His website is www.sce.carleton.ca/faculty/wainer.

XINGLIANG LI is a Developer, Team Leader at Ericsson. He received his PhD is in Telecommunications. His expertise is in algorithm development for 5G and LTE systems. His research is focused on performance verification, channel modeling, and beam management. His email is xingliang.li@ericsson.com.

GUOQIANG XUE is a manager of network system verification at Ericsson. His PhD is in Electrical Engineering. His team is specialized in 5G performance and capacity verification. Guoqiang is an expert in MIMO and beamforming testing and verification. He has filed many IPRs on the topics of MIMO testing and CDMA data transmission. His email is guoqiang.xue@ericsson.com.