Predicting Medication Prescription for Psychosis with Machine Learning Algorithms

EGEMEN SAHIN*, HAILEY FOX*, and UTTAM RAO*, Dartmouth College, USA

Additional Key Words and Phrases: datasets, neural networks, medical artificial intelligence, predicting prescription

ACM Reference Format:
Egemen Sahin, Hailey Fox, and Uttam Rao. 2022. Predicting Medication Prescription for Psychosis with Machine Learning Algorithms. *ACM Trans. Graph.* 37, 4, Article 111 (May 2022), 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Clinical decision support systems (CDSS) are computer programs designed to improve the quality of healthcare by assisting in medical decisions using electronic health records (EHR). As EHR data has grown, these technologies have become more accurate and efficient [18]. CDSS are used for a variety of different tasks including constructing clinical pathways, determining treatment plans, and predicting diagnoses. One manner in which CDSS are used is for learning doctor medicine prediction patterns. By training models on patient data and the medications that were prescribed, these technologies may be able to predict the best medication plan for a patient based on their health records.

One field in which this type of technology is especially useful is in the treatment of patients suffering from psychosis, a particularly difficult syndrome to treat. Psychosis is a clinical syndrome which is composed of a variety of symptoms that overall contribute to a disconnection from reality. Some symptoms include delusions, hallucinations, incoherent speech and confusion. Psychosis can be caused by a psychiatric illness such as schizophrenia or by another health condition. The severity and symptoms of psychosis vary greatly among patients, and thus treatment must be very personalized for each individual [5]. Current treatment for psychosis most often begins with the use of antipsychotic agents which are very effective against the more serious symptoms of psychosis [17]. Determining the best medication to prescribe has proven challenging as different patients may respond to the same medication very differently. Additionally, many medications can cause side effects which worsen the symptoms of psychosis [4]. Making sure that patients are given the correct medication is crucial for their treatment.

While it is very important that patients with psychosis are treated with the correct medication, very little research has gone into creating methods for psychosis medication prediction. Instead, much of the research has been focused on creating technologies for predicting psychosis diagnoses, monitoring the treatment plan for a patient with psychosis, or monitoring the symptoms of psychosis. We aim to address this gap in the space to create machine learning models to predict the medication that will be prescribed to patients using data collected during hospitalization. The models output a ranked prediction of medications that a doctor will prescribe. This not only could reduce the number of medications for a doctor to consider, but also detect potentially abnormal medication orders to prevent errors and reduce order time.

The current practice is for a clinician to determine which medication, or medications, they should prescribe to a patient based off of the individual's specific symptoms, medical history, observations made during physical examinations, and a wide variety of laboratory tests. Often, clinicians will prescribe multiple medications for a patient, a concept called polypharmacy. Psychiatric polypharmacy is the use of multiple medications for the treatment of a specific psychiatric medical condition. Evidence has shown that treating patients with psychosis with multiple medications can improve their outcomes [9]. We decided that instead of our algorithms predicting only one medication that will be prescribed to the patient, the algorithms should instead predict the probability of each medication being prescribed to the patient.

1.1 Objective

This research leverages a dataset containing information on psychosis patients collected by Hurd et al. [6] to train existing machine learning classifier architectures to predict how likely a patient is to be prescribed a variety of different medications. We train seven different classifiers that predict the probability that each medication will be prescribed to a patient at risk of psychosis upon their discharge. We aim to extract important features, and find trends that are useful in psychosis medication prediction. In order to do this, we train our algorithms to assign probabilities for each medication class in our data set, and then output all the medications which have a probability higher than a specified threshold.

A key contribution of our work is that there have been few studies aimed at predicting medication specifically for psychosis, and therefore this would be one of the first of its kind at this task. The possibility of predicting medication through patient data available at the time of admission and from the blood draws opens many possibilities for further research into utilizing data science methods in clinical environments to have a positive effect on patients or prevent side-effects and negative reactions from medication. Another benefit of our algorithm is that it could be helpful for newer doctors to see recommended medications based on the medications prescribed by more experienced doctors made for other patients.

*All authors contributed equally to this research.

Authors' address: Egemen Sahin, egemen.sahin.23@dartmouth.edu; Hailey Fox, hailey.p.fox.22@dartmouth.edu; Uttam Rao, uttam.rao.gr@dartmouth.edu, Dartmouth College, Hanover, New Hampshire, USA, 03755.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
0730-0301/2022/5-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

Many studies have focused on building algorithms for constructing treatment plans for patients suffering from psychosis. Our review of research revealed a gap in this study space: the lack of psychosis prescription predicting algorithms.

2.1 Medication prediction using EHR data

Although few are psychosis specific, there are several existing studies investigating using machine learning to predict patient medication based on patient data collected during hospitalization. For example, Rough et al. built models to predict patient-specific medication based on EHR data. Before their study, previous medication prediction research had generally been performed within a small population, with irregular timings, or by grouping medications into broad categories [12]. In this study by Rough et al., data on over three million medication orders (990 different types of medication) were used to train a deep learning sequence model (LSTM) as well as a logistic regression model. 55 percent of medications ordered by physicians were ranked in the top 10 predicted medicines sets of the sequence model (49 percent for logistic regression) and 93 percent of these sets included at least one medication that was prescribed [12]. Their results show that it is possible to predict patient medication based on EHR data.

In the Rough et al. study [12], the models do not aim to predict what medication a physician should prescribe (i.e what medications may have the best outcomes). Predicting the medication that results in the best outcome would require data collected after a patient has been discharged from the hospital such as patients' reactions to prescribed medications and the effectiveness of medications (which would be treated as the ground truth when training a model). Instead, the models in the Rough et al. study are built to predict what medication a physician actually will prescribe based on data collected during hospitalization. These predictions could be used by a doctor later on to guide their decision making based on prescriptions given by doctors on previous cases as well as detect abnormal medication orders to reduce errors. The goals of our project are similar to this, but focused specifically on the case of psychosis medication where there is a gap in the space.

2.2 Using patient data for medical predictions

Researchers at the University of Groningen created a Routine Outcome Monitoring (ROM) based computerized clinical decision aid Treatment-E-Assist (TREAT) for treating psychotic disorders [16]. This technology was created to monitor the effectiveness of a treatment plan for patients by summarising patient's unmet needs as well as providing evidence-based treatment recommendations for how to adjust the plan to make the patient more satisfied. They found that while clinicians did not believe it would likely improve patient outcomes, the technology was useful for their daily clinical practice and that they would like to use it in the future. While this method was not intended for medication prediction, these results do indicate that clinicians have found similar technology useful.

A study by Rajkomar et. al [11] shows that deep learning models can predict a patient's discharge diagnoses along with many other pieces of clinical information with a high accuracy, outperforming traditional predictive models in all cases. Their AI was trained on a dataset of three million medication records from 100,000 cases. Although this paper was not focused on the prescription of medication for patients that were at risk of psychosis, the findings from the study suggest that neural networks trained on patient data can be used as accurate and scalable predictive models.

There have also been studies which have developed algorithms aimed at finding prescription errors made by clinicians [7, 13]. A general trend from the conclusions of these studies showed that these algorithms were optimal for targeting high-risk patients whose reactions to medication are not as predictable. However, the objective of these studies are different from ours in that they do not predict the medication, but instead assess whether the medication prescribed by a clinician is at risk of being wrongly prescribed.

Machine learning has also been used to predict patients' prescriptions at an earlier stage of the patients' stays in clinics, in order to reduce pharmacy waiting times. Random forests models gave the highest accuracy at 97.22 percent for this task [3]. Their dataset included one million instances with 11 attributes each, and they predicted which of 161 drug codes would be the prescription for the patient. This paper showed that early prediction of prescription could be used to speed up the medicine-retrieval process when a patient is dismissed from a clinical environment.

3 METHODOLOGY

3.1 Dataset Description

Our dataset includes information on 119 patients with psychosis from the psychiatric unit of Mount Sinai Hospital in New York City [6]. Patients were diagnosed for acute psychosis by a clinician using the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) criteria and were approached for data collection within a week of hospitalization [1]. Patients were excluded for lack of fluency in English as well as history of various disorders or surgeries(traumatic brain injury, active seizure disorder, brain surgery, dementia, brain masses, etc.).

Three trained psychiatrists assessed the severity of psychosis in each patient using the Positive and Negative Syndrome Scale (PANSS) scale, which includes 30 different items to assess symptoms of schizophrenia [10]. There are a few cases where scale items are missing in the dataset. The dataset also includes data regarding the hospital course, medical history and comorbidity of patients. After patient discharge, additional data from patients' labs such as BMI, temperature, sleep length, blood count, and medication were added to the dataset. Data from blood and urine samples were collected from patients. These samples were mainly used to assess the presence of synthetic cannabinoids, THC, and other drugs. Data for some of these categories are missing for some patients due to patient refusal. In these cases a self reported drug history feature was substituted in the dataset. Lastly, data on inflammatory markers was collected for each patient. Analytes were measured for the following markers: IFN-γ, IL-12, IL-6, IL-2, IL-1β, TNF-α, IL-10, IL-21, IL-8, C-reactive protein, and soluble IL-2 receptors. Data regarding the batching of the samples is also included in the dataset. Overall, the dataset includes 108 features for each patient.

3.2 Exploratory Data Analysis

The dataset included 12 features indicating whether a certain medication was prescribed to a patient upon discharge. Predicting these features is the objective of this study. There needs to be sufficient data for each medication to train a model for prediction, which can be determined by counting how many patients had been prescribed a specific medication upon discharge for each medication.

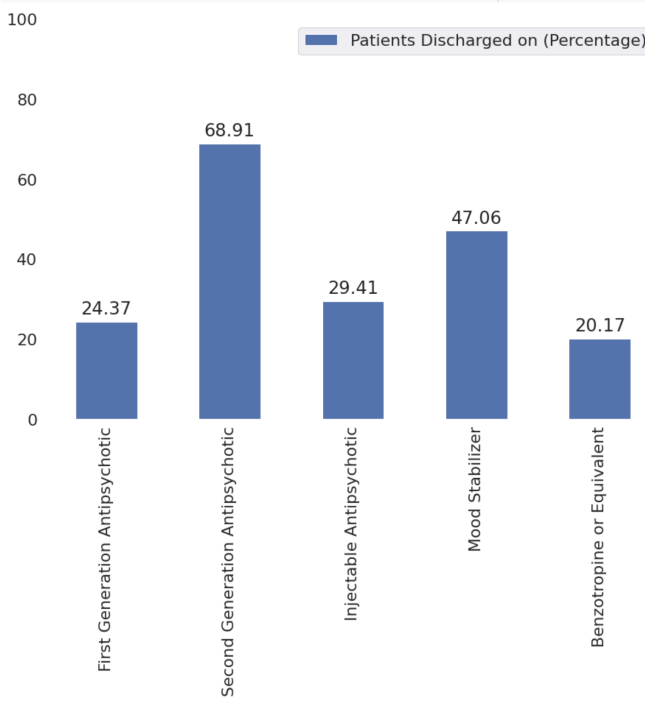

Fig. 1. Percentage of patients vs. medication type

Figure 1 shows the percentage of patients discharged on each type of medication. Since our dataset is relatively small and only includes information on 119 patients, any medications that were prescribed to less than 20 percent of patients were dropped from consideration. These medications were Clozapine, Antidepressant, Anxioytic, Medication to Prevent Relapse, and Panolol for Akathisia. We additionally dropped the "Any Antisypyschotic" column as it was just the union of the other antipsychotic types. We also dropped the "Benzotropine or Propanolol" column as it seemed to be a duplicate of the other benzotropine column, leaving us with a total of 5 medications to be predicted. We then counted how many medications were given to each patient to see if there were any specific groupings of medications (e.g. if most patients discharged on an injectable antipsychotic were also given benzotropine).

Figure 2, which counts patients by number of medications prescribed, shows that a majority of patients are prescribed more than one medication on discharge. Among these patients,60% were given exactly two medications, which was the highest percentage among the number of medications prescribed. There are no significant patterns or groupings within the medications given to patients. We

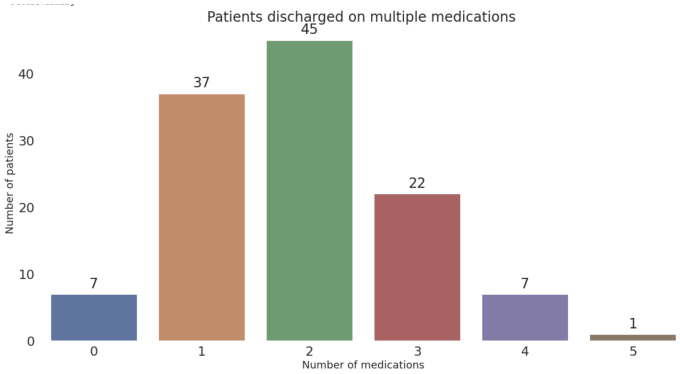

Fig. 2. Number of patients vs. number of medications

plotted the correlation between both numerical and categorical input features against the 5 medication labels. We were not able to discern any relationships between features and any of the labels, except "Psychotic disorder dx at discharge" and "neutrophils" which seemed to have some effect on the prescribed medication.

3.3 Data Pre-Processing

The dataset in this study includes 108 features, 12 of which relate to the prescribed medication (i.e., variable to be predicted). In addition, 39 other variables represent features like ID, Batch Number, etc. which are less relevant for the prediction task at hand. This yielded 57 features that were considered for training a model to predict the medication prescribed. After determining which features of the dataset we wanted to use as the input for our method, we then determined whether there was any missing data for these features. Any missing values (less than 2 percent) that we found for categorical features were imputed with the most frequent value. Categorical features included, for example, the "Group_Toxicology" feature which had values such as "Cannabinoid-Positive" or "Cannabinoid-Negative" instead of numerical values. For numerical features, any missing values were imputed with the median. We then used one hot encoding for all categorical features and standard scaling for all numerical features. There were 119 patients in our dataset, and after our pre-processing, there were 57 features for each patient. We then split the patients and their corresponding 57 features into training and test sets (80:20).

3.4 Multilabel Classification Strategies

Multilabel classification is a generalization of multiclass classification in which an input sample can be assigned multiple labels. Multilabel problems are generally approached in two ways: problem transformation and algorithm adaption [15].

Problem transformation refers to transforming a multiple label problem into multiple single label problems (binary or multiclass classification). The most common approach is called binary relevance, which is essentially training a binary classifier for each label and then combining the result to get a multilabel output. Binary relevance could use a one vs. one or a one vs. rest scheme with any binary classification method (e.g. logistic regression, SVM, simple

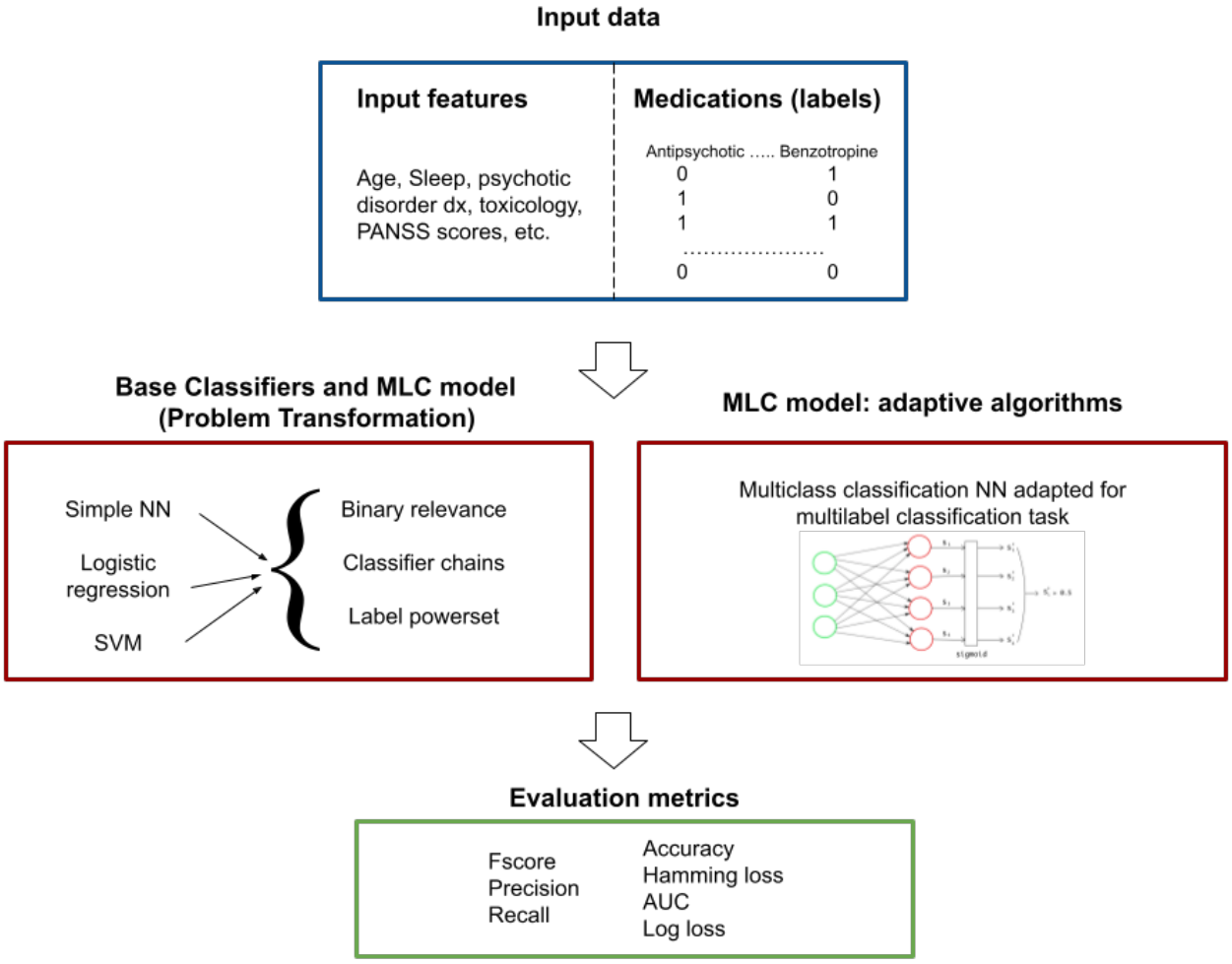

Fig. 3. Flowchart of methodology

MLP, etc.). One of the limitations of the binary relevance method is that it ignores any potential dependencies between labels. Unlike binary relevance, classifier chains consider potential correlations between labels by predicting labels sequentially and using the output of all previous classifiers as input into later classifiers. Algorithm adaption refers to extending specific models to work for multilabeling tasks without transforming the problem. These include neural networks (sometimes with special back propagation schemes), extensions of KNNs such as ML-kNN, and many others.

For our study, we trained a variety of algorithms to predict what medications were prescribed to a patient. We then compare these algorithms to determine the best performing one. We used both problem transformation and adaptive algorithms. We try a one vs. rest binary relevance with logistic regression and with a simple neural network. In addition, we also try classifier chains and label powersets. We also use a separate neural network that can directly handle multilabel classification tasks. All methods were implemented using scikitlearn and scikitmultilearn, which is a library used specifically for tasks such as multilabel classification. Tensorflow was used to build the neural networks.

3.5 Binary Relevance: One Vs. Rest

In order to use binary classification algorithms to solve our multi-classification problem, we utilized the one-vs-rest binary relevance multi-label classification method. Generally, this algorithm splits the multi-label classification dataset into multiple binary classification datasets and then fits a binary classification model on each. Overall, the multi-label classification is split into one binary classification problem per class.

3.5.1 Logistic Regression. As our dependent variable, whether or not a certain prescription was prescribed, is categorical, we decided to try logistic regression as the binary classification method utilized by our One-vs-Rest classifier. We utilized the sklearn OneVsRestClassifier package in order to do our processing.

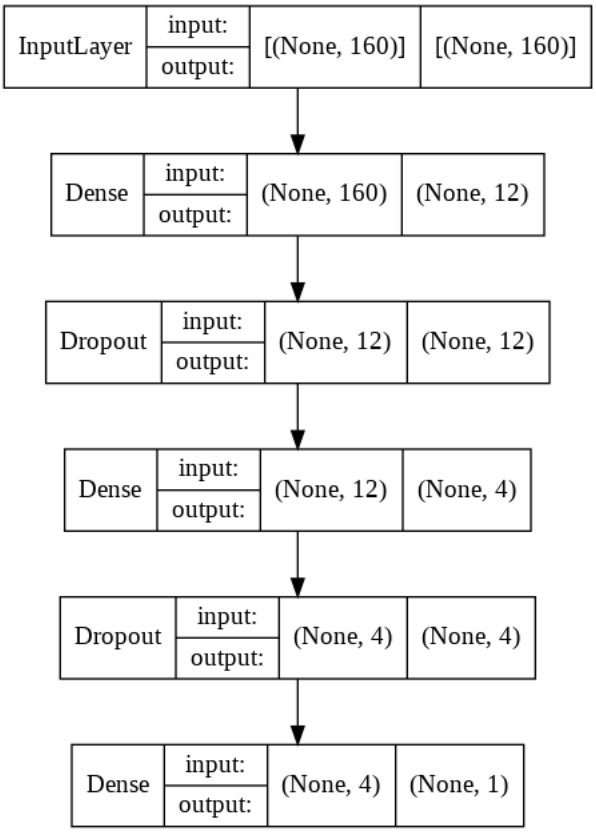

Fig. 4. Artificial Neural Network layers

3.5.2 Artificial Neural Network. We also used a small neural network as our base classifier for binary relevance. Artificial neural networks (ANN) are powerful data science tools. They are comprised of layers that contain perceptrons which receive input from the previous layer which is either another layer of perceptrons or the input. The perceptron equation (1) is applied to the input.

$$y = \phi(\sum_{1}^{n} W_i \cdot x_i) \tag{1}$$

ϕ represents the activation function, x represents the input and W represents the layer's weight. The ReLU (Rectified Linear Unit) activation function (2) was used in this ANN [2].

$$y = \max(0, x) \tag{2}$$

Due to the low amount of data samples in our dataset, there is the risk of the network memorizing the dataset which would lead to inaccurate testing predictions. To prevent the issue of overfitting, dropout layers are used. Dropout [14] is a regularization technique that prevents a percentage of neurons from being used in each training session.

The binary cross-entropy loss function was compiled on the ANN and Adam [8] was the optimizer. Parameter optimization was done on the model was done by experimenting with batch sizes of 3, 5 and 10, epoch numbers of 6, 30, 50, and 100, an ANN with or without a second Dense(4) layer, as well as dropout rates of 10%, 20% and 30%. The experimentations showed that the optimal model had a Batch size of 5 and an epoch number of 6 and consisted of one input, two hidden layers each with a 20% dropout rate and the output layer with 160, 12, 4 neurons and 1 neuron respectively, as illustrated in Fig. (4).

3.6 Classifier Chains, Label Powersets, and Adapted NN

We also implemented classifier chains and label powersets with both logistic regression and support vector machines. Additionally, another neural network was adapted to handle multilabel classification tasks directly, which the NN in Figure 4 cannot.

3.7 Evaluation Metrics

We use several metrics to evaluate the performance of the model including accuracy, precision, f1 score, recall, hamming loss, and AUC. For mutlilabel classification tasks, accuracy is a useful metric but was not as useful for our method as our dataset was imbalanced as seen in Figure 1, with for instance the "Any Antipsychotic" medication being prescribed to 90.76% of the patients while the "Benzotropine or Equivalent" medication was only prescribed to 20.17%. Hamming loss, which is the fraction on incorrectly predicted labels, is a more useful metric for our method because it allows us to detect not just if one medication was correctly predicted to be prescribed for a patient, but if each of the medications was correctly predicted to be prescribed or not for the patient. This is important as our model did not purely predict whether or not a medication was prescribed to a patient but instead what combination of medications was prescribed.

Due to the one hot encoding of the many categorical features, the input data is sparse. For some models, hamming loss is less sensitive when the training data is sparse and may not capture the performance of the model well. In these cases, log loss, which measures how close the prediction probability is to the true label, is a better metric to use. However, in the case of unbalanced frequency of labels such as in this study, log loss is very difficult to interpret. In this study, we use hamming loss as our main evaluation metric and consider log loss only to determine if a model is making a "dumb" prediction solely based on the prevalence of a label. Unlike with hamming loss, sklearn's built in log loss function does not support multilabel input so we develop our own.

4 RESULTS

4.1 Binary Relevance: ANN Results

Figure 5 shows the confusion matrices for binary relevance with the ANN as the base classifier. The predictions for the first generation antipsychotic, second generation antipsychotic, injectable antipsychotic, mood stabilizer and benzotropine medications had accuracies of 70.8%, 83.3%, 62.5%, 54.2% and 70.8% with F1 scores of 22.2%, 89.5%, 0%, 47.5% and 0% respectively. The main drawback of the ANNs was the false negative predictions seen in the confusion matrices. Most algorithms predicted negatives even when the test dataset contained no false outputs, which is a limitation that arose

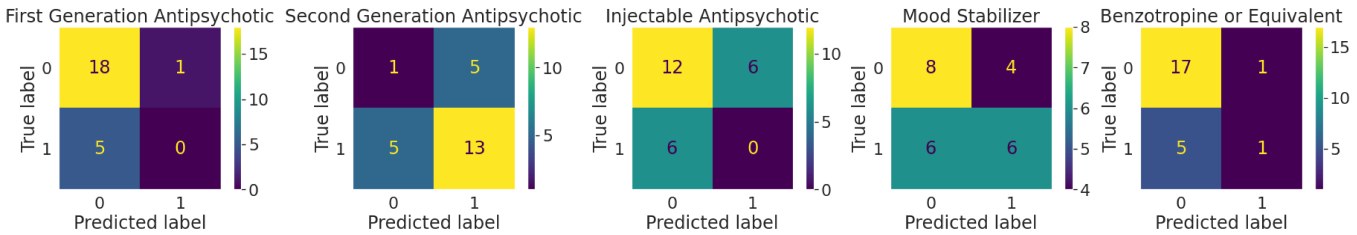

Fig. 5. Binary Relevance ANN: Confusion Matrices for Each Label

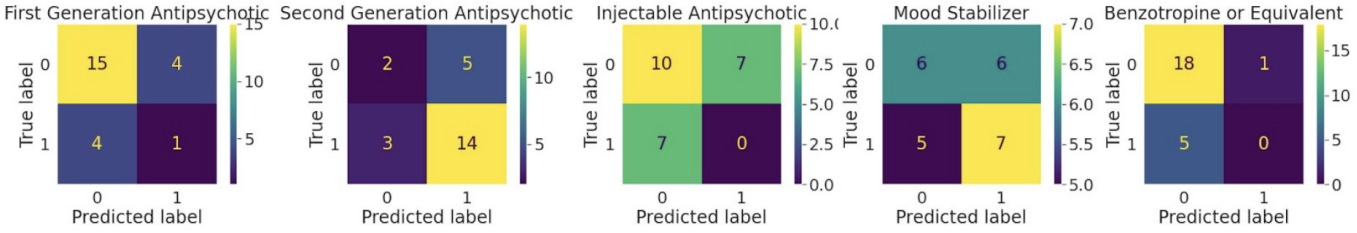

Fig. 6. Binary Relevance Logistic Regression: Confusion Matrices for Each Label

from the size of our dataset. This was reflected in the F1 scores which were either 0 or close to 0.

4.2 Binary Relevance: Logistic Regression Results

Figure 6 shows the confusion matrices for binary relevance with logistic regression as the base classifier. The predictions for the first generation antipsychotic, second generation antipsychotic, injectable antipsychotic, mood stabilizer, and benzotropine medications had accuracies of 66.7%, 66.7%, 41.7%, 54.2%, and 75.0% respectively.

4.3 Classifier Chain: SVM Results

Figures 7 and 8 show the classification report and the ROC curves for the classifier chain with support vector machines as the base classifier. The hamming loss for this model was 0.32. As figure 7 shows, this model only made any accurate predictions for "Second Generation Antipsychotic" and "Mood Stabilizer" and always predicted that no other medications were prescribed. Not coincidentally, the precisions of these two medications it did predict where 0.65 and 0.42 which are very similar to the percentage of patients prescribed those medications (see figure 1). In other words, this model seems to be only making predictions for the two most commonly prescribed medications and with a probability similar to their proportions in the training dataset.

4.4 Classifier Chain: Logistic Regression Results

Figures 9 and 10 show the classification report and the ROC curves for the classifier chain with logistic regression as the base classifier. The classifier chain with logistic regression had a hamming loss of 0.31 and an AUC of 0.67. It was very good at accurately predicting "Second Generation Antipsychotics" with a precision of 0.80, a recall of 0.84, and an f1 score of 0.82. Overall, the average precision and recall scores were about 0.6 across runs for all labels. This model

	precision	recall	f1-score
First Generation Antipsychotic	0.00	0.00	0.00
Second Generation Antipsychotic	0.65	1.00	0.78
Injectable Antipsychotic	0.00	0.00	0.00
Mood Stabilizer	0.42	0.22	0.29
Benzotropine or Equivalent	0.00	0.00	0.00
micro avg	0.60	0.40	0.48
macro avg	0.21	0.24	0.21
weighted avg	0.33	0.40	0.34
samples avg	0.58	0.44	0.47

Fig. 7. Classification Report for Classifier Chain Algorithm (SVM)

performed much better for every medication than the classifier chain with SVM as a base classifier.

4.5 Label Powerset: SVM Results

Figures 11 and 12 show the classification report and the ROC curves for the label powerset with SVM as the base classifier. The label powerset with logistic regression had a hamming loss of 0.38. The results for this model are neither good nor useful as it predicted that every patient in the test set was only given a "Second Generation Antipsychotic."

4.6 Label Powerset: Logistic Regression Results

Figures 11 and 12 show the classification report and the ROC curves for the label powerset with logistic regression as the base classifier. The label powerset with logistic regression had a hamming loss of 0.29 and an AUC of 0.64. This model performs similarly to the classifier chain using logistic regression.

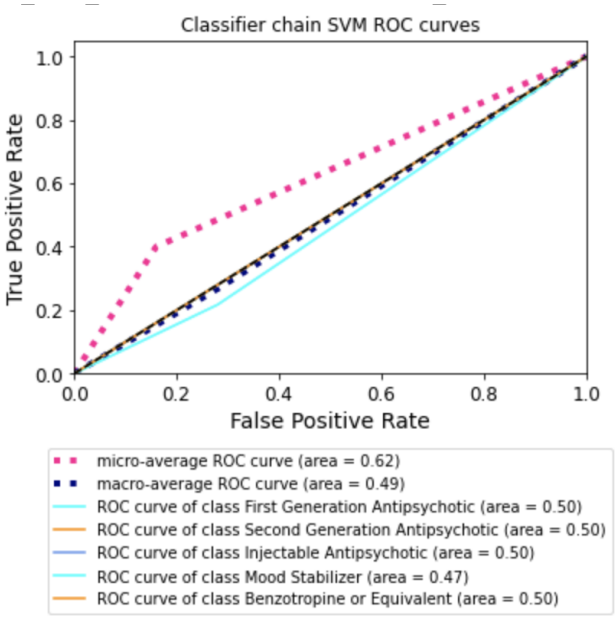

Fig. 8. ROC curves for Classifier Chain Algorithm (SVM)

	precision	recall	f1-score
First Generation Antipsychotic	0.33	0.20	0.25
Second Generation Antipsychotic	0.80	0.84	0.82
Injectable Antipsychotic	0.67	0.20	0.31
Mood Stabilizer	1.00	0.44	0.61
Benzotropine or Equivalent	0.50	0.14	0.22
micro avg	0.77	0.47	0.59
macro avg	0.66	0.36	0.44
weighted avg	0.75	0.47	0.55
samples avg	0.72	0.53	0.59

Fig. 9. Classification Report for Classifier Chain Algorithm (LR)

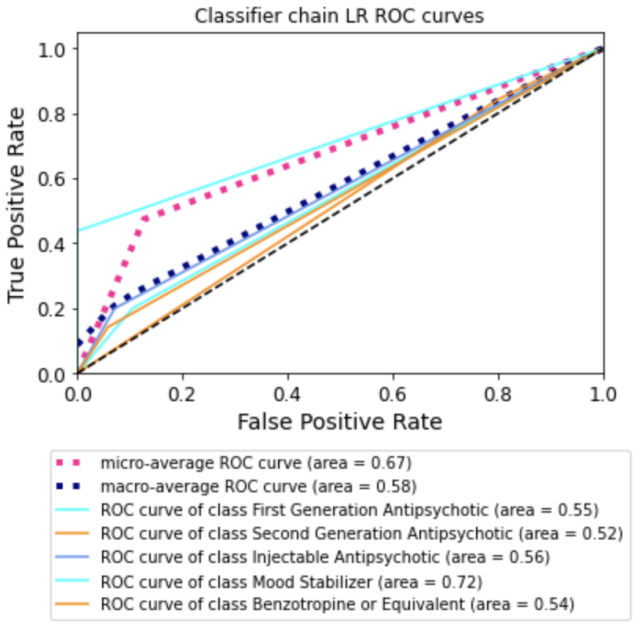

Fig. 10. ROC curves for Classifier Chain Algorithm (LR)

	precision	recall	f1-score
First Generation Antipsychotic	0.00	0.00	0.00
Second Generation Antipsychotic	0.79	1.00	0.88
Injectable Antipsychotic	0.00	0.00	0.00
Mood Stabilizer	0.00	0.00	0.00
Benzotropine or Equivalent	0.00	0.00	0.00
micro avg	0.79	0.33	0.47
macro avg	0.16	0.20	0.18
weighted avg	0.26	0.33	0.29
samples avg	0.79	0.41	0.52

Fig. 11. Classification Report for Label Powerset (SVM)

4.7 Adapted Algorithm NN

Figures 15 and 16 show the classification report and the ROC curves for the label powerset with logistic regression as the base classifier. The adapted algorithm NN has an AUC of 0.6 and a hamming loss of 0.34.

5 DISCUSSION

5.1 Limitations

Our study has several limitations which affected the perfomance of the models. First, the number of patients in the dataset is small. Despite removing medications that were prescribed to less than 20 percent of patients, the prevalence of many of the medications is still small and imbalanced. Second, the dataset includes a lot of categorical features which required preprocessing using one hot encoding. This leads to a lot of sparsity in the training data which makes it difficult for models to learn decision boundaries. Third, the dataset was originally collected for a slightly different purposes (investigating the effect of cannabinoid use on psychosis severity in hospitalized patients). Although the features present in the dataset seem to be suitable to make medication predictions, this may not be the case. These limitations led to the mediocre performance of our models. Machine learning models, especially neural networks, need a lot of training and test data available.

Despite these limitations some of our models showed promise. Using the log loss metric we determined if the models were simply making "dumb" guesses for medication predictions solely based on the prevalence of medication labels in the dataset. The results for the two SVM based models show that they were making these "dumb" guesses. Except for these two SVM based models, all of the models did better than "dumb" guesses, though the prevalence of the medications still had a clear effect.

5.2 Comparing models

Table 1 compares the performance of the various models tested in this study using hamming loss. Disregarding the SVM based models

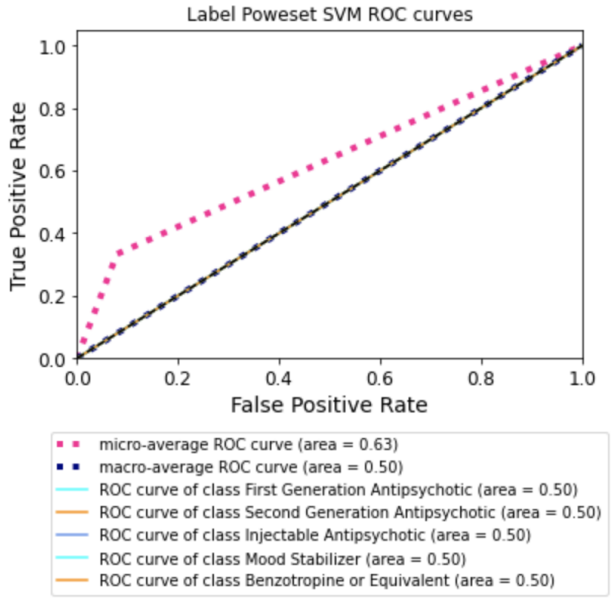

Fig. 12. ROC curves for Label Powerset (SVM)

	precision	recall	f1-score
First Generation Antipsychotic	0.25	0.20	0.22
Second Generation Antipsychotic	0.79	0.79	0.79
Injectable Antipsychotic	0.50	0.20	0.29
Mood Stabilizer	0.88	0.44	0.58
Benzotropine or Equivalent	0.50	0.29	0.36
micro avg	0.69	0.47	0.56
macro avg	0.58	0.38	0.45
weighted avg	0.68	0.47	0.54
samples avg	0.68	0.51	0.55

Fig. 13. Classification Report for Label Powerset (LR)

(see limitations section) the logistic regression based label powerset and classifier chain models performed the best.

Comparison of Models	
Model	Hamming Loss
Classifier Chain (SVM)	0.32
Classifier Chain (LR)	0.31
Label Powerset (SVM)	0.38
Label Powerset (LR)	0.29
Adapted NN	0.34
Binary Relevance (ANN)	0.32 (average)
Binary Relevance (LR)	0.39 (average)

Table 1. Hamming loss by model

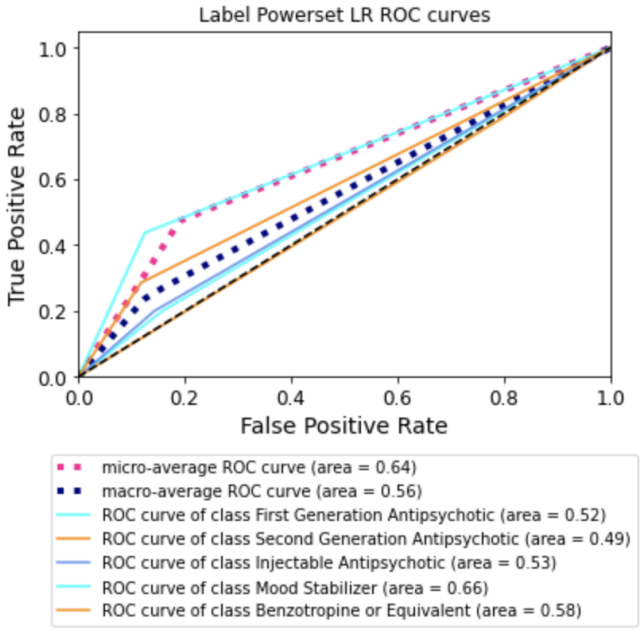

Fig. 14. ROC curves for Label Powerset (LR)

	precision	recall	f1-score
First Generation Antipsychotic	0.43	0.25	0.32
Second Generation Antipsychotic	0.77	0.55	0.64
Injectable Antipsychotic	0.30	0.79	0.43
Mood Stabilizer	0.44	0.65	0.53
Benzotropine or Equivalent	0.75	0.30	0.43
micro avg	0.47	0.54	0.51
macro avg	0.54	0.51	0.47
weighted avg	0.57	0.54	0.51
samples avg	0.44	0.50	0.44

Fig. 15. Classification Report for the Adapted NN

5.3 Further work

The performance of classifier chains are dependent on the order in which labels are considered, but we did not test multiple orderings. Further work could test different orders of labels to see if it impacts performance.

For all of the models described in this paper, we can get probabilities for each class and rank the medications based on those probabilities. A metric which might provide even better analysis of than Hamming loss would be one which measures the accuracy of the medication ranked highest actually being prescribed. For example, predicting the top 2 prescribed medications correctly and getting the other three wrong is more important than getting the other three right and the top 2 wrong (but neither accuracy or hamming loss would reflect this). This is similar to methods used by [12]. For this study, our preliminary results did not warrant us investigating the results for this metric.

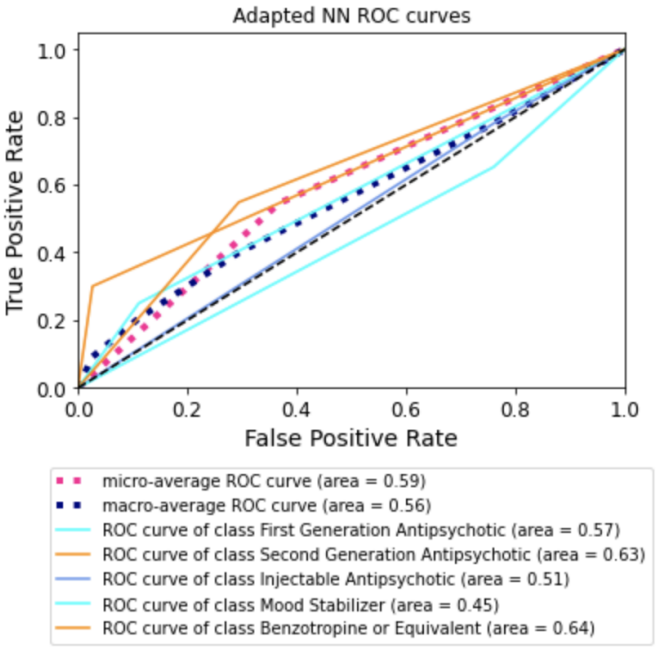

Fig. 16. ROC curves for the Adapted NN

An interesting insight from the Rough et al. study [12] is that long short-term memory networks (LSTMs, which can process sequenced data) seem to work quite well for the type of prediction task in our study. Although our dataset was not originally collected and used with analyzing sequential data in mind, it may be possible to split the dataset into parts, with the first part being the physiological data of the patient which is gathered at the time of admission, and the additional parts being the data that was gathered from the lab tests that are available 1 to 3 days after the admission (based on the procedure used in the extraction of our dataset). By transforming the data into a sequenced dataset, real-life scenarios where the insights from the laboratory tests are not immediately available when the patient is admitted could be better simulated. An advantage of organizing our data this way is that the LSTM deep learning algorithm can be used. Alternatively, a future work could use a better dataset with sequential data.

6 CONCLUSION

The objective of our study was to create a method for predicting how likely a patient is to be prescribed a variety of different medications. We tested 7 different models that predict the probability that each (of five) medications will be prescribed to a patient at risk of psychosis upon their discharge. Due to several limitations, primarily related to the size of the dataset, our trained models did not have great performance. However, future work could continue this research using similar methods but with a more suitable dataset.

REFERENCES

[1] 1970. DSM 5 Diagnostic and Statistical Manual of Mental Disorders. (1 January 1970). https://pesquisa.bvsalud.org/portal/resource/pt/psa-52826

[2] Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). (2018). https://doi.org/10.48550/ARXIV.1803.08375

[3] Alshammari R. Alhorishi N, Almeziny M. 2021. Using Machine Learning to Predict Early Preparation of Pharmacy Prescriptions at PSMMC - a Comparison of Four Machine Learning Algorithms. *Acta Inform Med* (2021).

[4] Graham C.L. Davey. 2014. Overprescribing Drugs to Treat Mental Health Problems. *Psychology Today* (30 January 2014). https://doi.org/us/blog/why-we-worry/201401/overprescribing-drugs-treat-mental-health-problems

[5] Zielasek J. Gaebel, W. 2015. Focus on Psychosis. *Dialogues Clin Neurosci* 17, 1 (2015). https://doi.org/10.31887/dcns.2015.17.1/wgaebel

[6] Yasmin Hurd, Anahita Bassir Nia, Madhu Mazumdar, Daniel Defrancisco, Manishkumar Patel, Seunghee Kim-Schulze, Nicolas Fernandez, Sharron Spriggs, Claire L. Gibson, Xiaobo Zhong, Amy Swift, and Charles Perkel. 2020. Data for: Cannabinoid use in psychotic patients impacts inflammatory levels and their association with psychosis severity. *Mendeley Data* V1 (2020). https://doi.org/10.17632/rb3xzfwxss.1

[7] Olivier Martin Xavier Dode Jean-Patrick Lajonchère Olivier Billuart Yvonnick Bézie Anne Buronfosse Jennifer Corny, Asok Rajkumar. 2020. A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. *Journal of the American Medical Informatics Association* 27, 11 (November 2020).

[8] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). https://doi.org/10.48550/ARXIV.1412.6980

[9] Kaira Gurvinder Shah Nilesh Shrivastava Amresh Kukreja, Sanjay. 2013. Polypharmacy in Psychiatry: A Review. *Mens Sana Monogr* (2013), 82–99. https://doi.org/doi:10.4103/0973-1229.104497

[10] Michael Obermeier. 2012. The positive and negative syndrome scale for schizophrenia: an established rating instrument in need of clarification. (2012).

[11] Oren E. Chen K. et al. Rajkomar, A. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Med* 1, 18 (2018). https://doi.org/10.1038/s41746-018-0029-1

[12] K. Rough, A. Dai, K. Zhang, Y. Xue, L. Vardoulakis, C. Cui, A. Butte, M. Howell, and A. Rajkomar. 2020. Predicting Inpatient Medication Orders From Electronic Health Record Data. *Clinical pharmacology and therapeutics* (2020).

[13] Volk LA Forsythe KJ-Myers S McGurrin M Williams DH Bates DW Schiff G Seoane-Vazquez E Rozenblum R, Rodriguez-Monguio R. 2020. Using a Machine Learning System to Identify and Prevent Medication Prescribing Errors: A Clinical and Cost Analysis Evaluation. *Jt Comm J Qual Patient Saf.* 46, 1 (January 2020), 3–10.

[14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[15] Piotr Szymański. 2018. Relevant Concepts in Multi-Label Classification. (2018). http://scikit.ml/concepts.html

[16] Roebroek-L.O. Lieburg E.J. et al. Tasma, M. 2018. The development and evaluation of a computerized decision aid for the treatment of psychotic disorders. *BMC Psychiatry* 18, 163 (2018). https://doi.org/10.1186/s12888-018-1750-7

[17] WHO. 2009. Pharmacological Treatment of Mental Disorders in Primary Health Care. *Geneva: World Health Organization* (2009).

[18] Mei-J. Xie G. Li X. Li-Z. Xu M. Xia, E. 2017. Learning Doctors' Medicine Prescription Pattern for Chronic Disease Treatment by Mining Electronic Health Records: A Multi-Task Learning Approach. *AMIA Annu Symp Proc* (2017).