Accuracy Enhancement of Auto-Diagnosis Models

Ting-Hsuan Chung
chung.tin@northeastern.edu
Khoury College of Computer Science,
Northeastern University

Jiaying Zheng
zheng.jiay@northeastern.edu
Khoury College of Computer Science,
Northeastern University

Zihan Zhao
zhao.zih@northeastern.edu
Khoury College of Computer Science,
Northeastern University

Abstract

Applying machine learning models to assist in diagnostics is a popular topic nowadays, with many different applications in the real world. However, how does the performance of these models vary among other approaches, and how can the excellent performance model be applied to real-world use? This paper addresses the challenge of enhancing the accuracy of pathology and differential diagnosis prediction by applying several machine-learning models to a large-scale synthetic dataset and then applying the best performance model in differential diagnosis prediction into a simple inquiry system to pretend the real-world use. The models include Logistic Regression, Random Forest, Weighted Ensemble, Neural Network, Modified Transformer, and LSTM. Through the work, the Random Forest model achieved the best performance in pathology prediction among these models. Meanwhile, the LSTM Model showed superior performance across almost all metrics in differential diagnosis, indicating its robustness in processing large, complex patient datasets. The Auto Diagnosis(AD) system, by applying LSTM, also achieved a better result comparing to previous researches. This project showed the potential of machine learning to develop diagnostic processes in healthcare, indicating a way to a more efficient and accurate auto diagnosis system.

1 Introduction

In the field of healthcare, the integration of machine learning technologies has played an important role(Haug & Drazen, 2023)(Miotto et al., 2018), particularly in the realm of disease diagnosis and management. This project focuses on the critical challenge of enhancing the accuracy and efficiency of pathology diagnosis and differential diagnostics and then applying the best performance model

of differential diagnostics prediction into a simple inquiry system to show the auto-diagnosis in real-world use. Our method utilizes a vast electronic health record (EHR) and patient-generated data dataset, applying diverse machine learning models to predict potential pathologies and create differential diagnoses and building a simple inquiry system for applying the outstanding performance of the differential diagnosis prediction model into an auto-diagnosis system. It showed the potential to develop the diagnostic process further, making it more precise, efficient, and less reliant on subjective human interpretation.

We aim to find more effective ML models for pathology and differential diagnosis prediction. Our group Employed a diverse array of models, including Logistic Regression, Random Forest, Weighted Ensemble for pathology prediction, and Neural Network, Modified Transformer, and LSTM for differential diagnosis. This study aims to push the boundaries of current medical diagnostic capabilities. The choice of these models was informed by their proven efficacy in handling large, complex datasets, which is crucial given the intricate nature of medical data and the need for high precision in diagnostics.

The contribution of this project is its potential to improve diagnostic processes within the healthcare system. By harnessing the power of Machine Learning, it enhanced the accuracy of diagnoses. Applying the superior model of predicting differential diagnosis into a simple inquiry system to build an auto-diagnosis system makes our study more like real-world use. To illustrate our study, our group provided a detailed exploration of our methodology, findings, and the implications of our work for the future of medical diagnostics.

2 Related Work

Machine learning has emerged as significant topic in healthcare sector recently. In (Huang et al., 2022), a transformer-Based model was used to conduct differential diagnosis of demyelinating diseases based on conventional Magnetic Resonance Imaging (MRI). In (Eftimie et al., 2022), they applied Random forest for the differential diagnosis of complex pathologies using image features from microscopic images of thyroid nodule capsules which highlights RF's proficiency in managing high-dimensional data and identifying diagnostic patterns.

However, above papers focused on using image to predict the disease rather than using Electronic Health Record(EHR) or evidence retrieved from doctor-patient interaction which appears in the clinical process.

Auto diagnosis in the clinical process is another emerging topic in healthcare area. Several academic paper has demonstrated its importance and provided different methods. Ensemble method is an approach used in auto diagnosis system. In (Sunena Rose & Sobhana, 2021), deep Learning, support vector machine (SVM), and decision tree were used to build an ensemble model as an automated system for disease diagnosis. (Kaushik et al., 2020) demonstrated the superiority of ensemble models, which integrate ARIMA, MLP, and LSTM predictions, which highlighted the potential for adopting similar ensemble approaches to enhance the accuracy and reliability of automated disease diagnosis systems.

LSTM model is a quite popular approach for diseases prediction. (Men et al., 2021) presents a deep learning framework employing LSTM networks enhanced with time-aware and attention-based mechanisms for multi-disease prediction from patient clinical records from a China's southeastern hospital.

DDXPlus(Fansi Tchango et al., 2022) is a innovatively created large dataset that incorporate differential diagnosis, pathology and multi choice of symptoms and antecedents. It allows a more nuanced and effective way for data collection which mimics real patient-doctor interaction. In this paper, methods mentioned in (Yuan & Yu, 2024)(AARLC) and (Luo et al., 2021)(BASD) are applied on the dataset to predict pathology and differential diagnosis. However, these approaches failed to achieve promising result in differential diagnosis prediction. Following by (Alam et al., 2023), a transformer model was used to the same dataset and reach higher metrics, while the inquiry system was not included in the model.

In this work, we used the DDXPlus dataset to perform model comparison in differential diagnosis. AARLC and BASD were used to perform comparison in auto-diagnosis system.

3 Problem Statement

In the rapidly evolving field of medical science, traditional methods of pathology diagnosis and differential diagnostics are facing new challenges due to the growing complexity and volume of medical data. Accurate medical diagnosis is critical for patient treatment and recovery, and manual diagnostic processes are often time-consuming and subject to subjective bias. Also, building an auto-diagnosis system for clients' use is crucial since the auto-diagnosis system can better simulate the real diagnosis and treatment. The problem is to use the patient data, which may include any related information like age, sex, symptoms, etc., to predict what pathology the patient is experiencing or to generate a differential diagnosis. It is a classification model that classifies a single or several possible pathologies. It is also a challenge to collect the patient's information from clinical inquiry. If we can build an accurate system, it can be used to assist doctors in making better and quicker diagnoses.

4 Dataset

The dataset covers the scope of the problem we are investigating. The dataset presents a large-scale synthetic dataset of roughly 1.3 million patients that includes a differential diagnosis, along with the ground truth pathology, symptoms and antecedents for each patient. Since our topic is developing a system to recommend doctors for possible diagnoses, the content of the dataset fits our goal of the project.

The data is in csv format and json format which we can use python to analyze. It contains these variables:

- `AGE`: the age of the synthesized patient.
- `SEX`: the sex of the synthesized patient.
- `PATHOLOGY`: name of the ground truth pathology that the synthesized patient is suffering from.
- `EVIDENCES`: list of evidences experienced by the patient. An evidence can either be binary, categorical or multi-choice. A categorical or multi-choice evidence is represented in the format '[evidence-name]_@_[evidence-value]' where ['evidence-name'] is the name of the evidence and ['evidence-value'] is a value from the 'possible-values' entry. A binary evidence is represented as '[evidence-name]'.
- `INITIAL_EVIDENCE`: the evidence provided by the patient to kick-start an interaction with an ASD/AD system. This is useful during model evaluation for a fair comparison of ASD/AD systems as they will all begin an interaction with a given patient from the same starting point. The initial evidence is randomly selected from the binary evidences found in the evidence list mentioned above (i.e., `EVIDENCES`) and it is part of this list.
- `DIFFERENTIAL_DIAGNOSIS`: The ground truth differential diagnosis for the patient. It is represented as a list of pairs of the form '[[patho_1, proba_1], [patho_2, proba_2], ...]' where 'patho_i' is the pathology name and 'proba_i' is its related probability.

The data is from the paper, here is one drawback of the dataset: *"Although the overall dataset construction process is reasonable, it has several shortcomings due to the fact that the dataset is completely synthetic, and completely relies on existing medical ontology and a rule-based software."*

4.1 Data Preprocessing Steps

For our purpose, we processed the dataset by following steps:

1. Ensured there are no outliers in `AGE` and `EVIDENCES`
2. Turn `EVIDENCES` from string to list
3. One-hot encode evidences
4. Turn column `SEX` into binary form

For predicting pathology, we also have to drop `DIFFERENTIAL_DIAGNOSIS` and `INITIAL_EVIDENCE` columns.

For predicting differential diagnosis, we need to drop the column `INITIAL_EVIDENCE`. Also, we split `DIFFERENTIAL_DIAGNOSIS` column into multiple columns, with each column named after a specific pathology and containing the probability values.

5 Methodology

For pathology categorization, we decided to choose from traditional supervised machine learning models. Because of the size of the dataset, traditional machine learning can achieve high accuracy without overfitting. Our dependent variable has 49 categories so we have to choose models that can categorize multiple categories. Considering all these above, we chose Logistic Regression, Random Forest and Weighted Ensemble. Because Logistic regression is interpretable and is efficient on large datasets. Random forest model is able to handle high dimensionality and robust to noise. Weighted Ensemble can improve overall accuracy and reduce overfitting.

For differential diagnosis prediction, we have to choose from models that can accurately predict multi-categories. Considering the size of the dataset, we have to choose models that are able to deal with large dataset. we first considered Random Forest but didn't select it due to the time consumed. At last, we chose Neural Network, Modified Transformer and LSTM. We chose Neural Network because it can handle non-linear relationships and large dataset, and is flexible in architecture. We chose Modified Transformer because we collectively accept the strength of Transformer model in (Alam et al., 2023). However, we believe that through modifying it, we can achieve better outcomes and make the model more logical. Besides, we also chose the LSTM model because of its exceptional ability to process and understand long sequences and complex data dependencies, which is crucial for the accurate prediction of differential diagnoses, especially in scenarios involving large datasets.

For building an Auto Diagnosis(AD) system by incorporating an inquiry system to the best model above, we introduced a simple inquiry system based on entropy.

The overall methodology is shown in Figure 1.

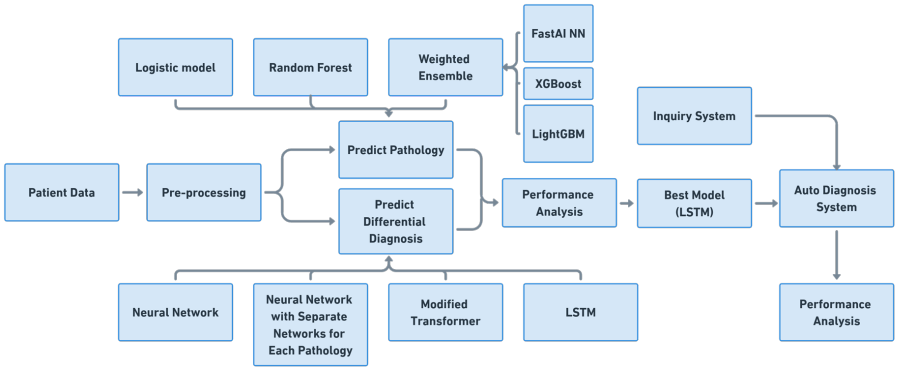

Figure 1: The overall methodology

5.1 Predicting Pathology

5.1.1 Logistic Regression

First, we predicted pathology outcomes using a Logistic Regression model, chosen for its suitability for classification tasks. During the training phase, we determined that the best parameters for handling this large dataset were the default settings, which include the 'lbfgs' solver and the standard number of iterations. We also employed L2 regularization, an inherent feature of scikit-learn's Logistic Regression, to counteract overfitting by penalizing large weight coefficients.

5.1.2 Random Forest

We also used Random Forest model to predict pathology. The Random Forest algorithm was chosen due to its proficiency in managing large datasets and its inherent mechanisms to prevent overfitting. We did not specify a batch size or number of epochs, as these are not applicable to Random Forest training. To optimize our model, we employed a grid search strategy with a 5-fold cross-validation approach. The best parameter we find is: max_depth=20, min_samples_leaf=2, min_samples_split=4, and random_state=0. We used cross-validation within the grid search process to enhance the model's generalizability. After the grid search, we further validated the model on a separate validation set to confirm the model's performance and to check for overfitting.

5.1.3 Weighted Ensemble

Different models may be better at capturing different aspects of the data. By using a weighted combination of models, an ensemble can leverage these diverse perspectives to make more informed predictions. Our weighted ensemble model as shown in Figure 2 is an ensemble model of Neural Networks, LightGBM, and XGBoost. The implementation was done by AutoGluon, in which we used 5-fold cross-validation for each submodel. We set presets as "medium_quality" time limit as 3600 seconds. All other hyperparameters are default in AutoGluon.

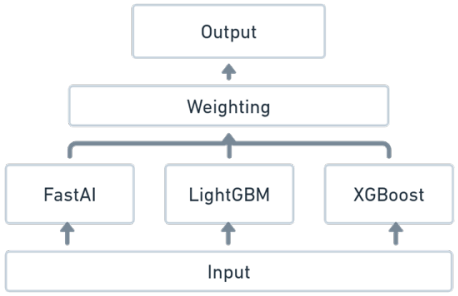

Figure 2: The weighted ensemble model structure

For validation, we used the validation set to make predictions with the previously trained models and calculated the metrics. The results indicated that there should not be an overfitting issue with our models.

To evaluate the performance of the above three models, we used the same test data and compared the prediction to the "pathology" column.

5.2 Predicting Differential Diagnosis

5.2.1 Neural Network

For this model, the architecture consisted of densely connected layers with LeakyReLU activation and dropout for regularization. We also employed batch normalization to stabilize and accelerate training. The model had input layer with a size corresponding to the feature space; dense layer with 1024 units, LeakyReLU activation, and Batch Normalization; dropout layer with a rate of 0.3; second Dense layer with 512 units, followed by LeakyReLU and Dropout; and an output layer with a sigmoid activation function corresponding to the number of diagnostic categories.

The model was trained using an Adam optimizer with a learning rate of 0.00005. The loss function was binary cross-entropy, suitable for multi-label classification tasks. The batch size was set at 128, and the model was trained for up to 200 epochs with early stopping implemented to avoid overfitting. The early stopping callback monitored the validation loss with a patience of 20 epochs and restored the best weights for the final model.

We used validation set to tune hyperparameters and prevent overfitting. The performance on the validation set guided adjustments in learning rate, layer size, dropout rate, and early stopping criteria. This iterative process ensured the model generalized well beyond the training data.

5.2.2 Neural Network with Separate Networks for Each Pathology

This model used Neural Network with FastAI backend by Autogluon to train each label(pathlogy) based on the values in ground true differential diagnosis as a regression task. Each label had its own network and took the previous outputs of networks as input. Thus, the prediction of a specific label can only be obtained if previous predictions have made. The brief structure of the model is shown in Figure 3. All hyperparameters are default in AutoGluon but a time limit of 210 seconds is set for each network.

5.2.3 Modified Transformer

We also implemented a transformer model for predicting differential diagnosis. The modified transformer model is basically modified from (Alam et al., 2023) which the transformer structure is from (Dong et al., 2021). We rebuilt the vocabulary in the model without replacing "_@_" with " ", which provided each choice in a categorical evidence as a token to the model. We also disabled

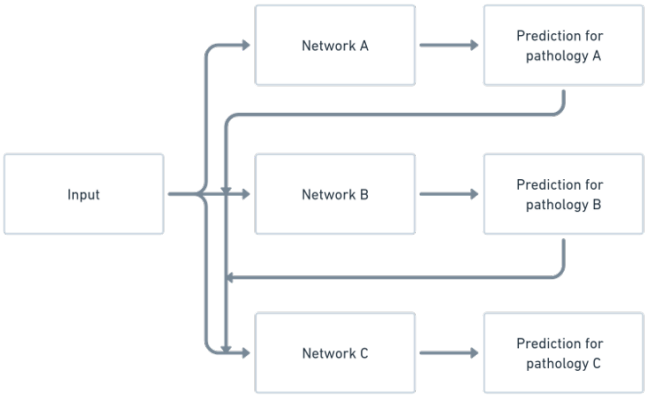

Figure 3: Neural Network with Separate Networks model structure. There are 49 networks like Network A. The previous outputs are provided to all the following networks as a input.

the positional embedding in the encoder since the sequence in the input evidence is not related to the differential diagnosis. We trained the model for 8 epochs. Rest of parameter remained the same as (Alam et al., 2023).

5.2.4 LSTM

Moreover, we implemented a Long Short-Term Memory(LSTM) model for predicting the differential diagnosis to avoid the vanishing gradient problem of traditional recurrent neural networks(RNN). The LSTM model is trained to classify sequential data over 40 epochs using a batch size of 64 for optimized gradient descent. We first converted the data into PyTorch tensors, subsequently loaded into a 'DataLoader' to enable efficient batch processing and shuffling. The model ,comprising two LSTM layers followed by two linear layers. Then an Adam optimizer with an initial learning rate of 0.001 updates the model's weights. The chosen loss function is the 'BCEWithLogitsLoss'. This loss function is particularly suited for binary classification tasks because it combines a sigmoid activation function with the binary cross-entropy loss in a single operation, streamlining the training process. The training loop includes a forward pass, loss computation, gradient rest, backpropagation, and a parameter update. Progress is monitored by logging the loss every 100 steps, ensuring transparency and control over the model's learning trajectory.

To evaluate the performance of the above three models, we converted the prediction of the model into a binary array and used the same test data with its differential diagnosis also as a binary array to compare with.

5.3 Building AD system

In order to build an AD system, we introduced a simple inquiry system to combine the LSTM model, which is the best one we have. Here is how the inquiry system works:

1. Find the question that has the largest entropy across the whole training dataset and make it as the first question.
2. According to the patient's answer, find the similar cases by applying Manhattan distance to all the cases in the training dataset.
3. Within the similar cases find the question that has largest entropy and make it as the next question
4. Repeat Step 2 by combining all previous answers and Step 3 until the question length reaches the setting.

For detail parameters, we set the number of questions to be 20. After three questions, the age of the patient was taken into consideration of similar cases. We chose not to take the sex of patient into consideration of similar cases since the result in EDA shows that the sex makes no difference to patient's pathology. However, it still a parameter to LSTM predictive model. Also, the age of the patient was assigned to 6 intervals which with the maximum distance of 5. In addition, the difference in each answer was considered a distance of two instead of one to mitigate the distance caused by the age. With the first patient's answer, 5% of most similar cases were consider to calculate the next question. After that, the following percentages of similar cases was calculated as:

$$percentage_{i+1} = \begin{cases} \lfloor \frac{percentage_i}{1.4^{\lfloor entropy/2 \rfloor + 1}} \rfloor & \text{when} \geq 0.01 \\ 0.01 & \text{otherwise} \end{cases}$$

The data processed by the inquiry system will consist of 0 for unknown, 1 for positive response, and -1 for negative response. After processing all train data by the inquiry system, we trained the LSTM model by the processed data. We used the same parameters with previous introduced LSTM model above.

6 Experimentation and Result

6.1 Experiment Setup

We aim to provide a better prediction of pathology and differential diagnostics. The primary objective of our project is to imitate and enhance the methodologies presented in existing research, with the methods mentioned above, thereby improving the accuracy of pathological predictions and differential diagnostics.

Our proposed models were trained and evaluated on a laptop with Intel Core i7-12650H CPU and GeForce RTX 3060 Laptop GPU. All models were written in Python. Further environment setup can be found in Github page.

6.2 Result

Besides evaluating between our models, we included comparing the performance of the models from (Fansi Tchango et al., 2022) and (Alam et al., 2023) (below we use the word "paper" to refer them).

For the pathology, the metrics chosen for the purpose are accuracy, F1 score, precision, recall, and balanced accuracy. We choose these metrics so that we can compare with the paper:

Accuracy:It is the ratio of correctly predicted observation to the total observations.

Precision: It indicates the proportion of positive identifications that were actually correct.

Recall: This metric gives us the proportion of actual positive cases that were correctly identified.

F1 Score: The F1 score is the harmonic mean of precision and recall. It is a useful measure when seeking a balance between precision and recall, because the class distribution in the data is uneven.

Balanced Accuracy: This is the only metric that is different from the original paper. We choose this because the dataset is imbalanced. It calculates the average of recall obtained on each class, ensuring that each class is equally important.

When predicting differential diagnosis, We also choose to use the same metrics with the paper:

IL (Interaction Length): The number of questions that will asked to the patient.

DDR (Disease Detection Rate): This metric measures the rate at which the model correctly identifies the presence of each disease. It is crucial for ensuring that the model reliably detects each possible condition.

DDP (Disease Discrimination Precision): DDP is the model's precision in distinguishing between different diseases. This is important in a differential diagnosis context, as it ensures that the model accurately identifies the specific disease from a range of possibilities.

DDF1 (Differential Diagnosis F1 Score): Similar to the traditional F1 score, DDF1 is the harmonic mean of precision and recall, but in the context of differential diagnosis. It provides a balance between the precision and recall across all predicted diseases.

GM (Geometric Mean): The geometric mean of sensitivity and specificity for each disease. This metric is useful for imbalanced datasets and provides an overall effectiveness measure of the model across all diseases.

The final model was evaluated on a separate test set to estimate its real-world performance.

For the three models shown in Table 1, weighted Ensemble and Random Forest model reached the highest accuracy of 99.72%. For accuracy of models in (Fansi Tchango et al., 2022) and (Alam et al., 2023), AARLC reached 99.21%,

Table 1: Comparison of Pathology Classification Methods

Method	**Accuracy (%)**	**Precision(%)**	**Recall(%)**	**F1 score**	**Balanced Accuracy(%)**
Logistic Regrssion	99.19	96.64	94.64	0.9498	94.64
Random Forest	**99.72**	**99.73**	**99.64**	**0.9967**	**99.63**
Weighted Ensemble	**99.72**	99.73	99.63	0.9966	99.62

BASD was 97.15%, and DDxT was 99.8%. The accuracy of our model is higher than the ones in DDXPlus: A New Dataset For Automatic Medical Diagnosis. Since predicting pathology can be a separate part and output in the system, we choose the weighted ensemble model output the prediction of pathology. The accuracy of 99.7% is used to calculate the GM in Table 2.

Table 2: Comparison of Differential Diagnosis Prediction Models

Method	**DDR**	**DDP**	**DDF1**	**GM**
DDxT	94.84	94.65	0.9472	96.45
Neural Network	90.55	**99.60**	0.9400	96.52
NN w Separate Network	99.32	90.89	0.9410	96.55
Modified Transformer	95.03	98.08	0.9629	97.58
LSTM	**98.28**	98.16	**0.9802**	**98.71**

Table 3: Comparison of AD System Performances

Method	**IL**	**DDR**	**DDP**	**DDF1**	**GM**
AARLC	25.75 (2.75)	69.53	**97.73**	0.7824	87.68
BASD	17.68 (0.88)	88.34	85.03	0.8369	90.03
AD-LSTM	20	**89.21**	88.81	**0.8862**	**92.44**

The result of predicting differential diagnosis is shown in Table 2. Among the models, NN w/ Separate Network achieved the highest DDR of 99.32 %, showing its ability to correctly identify the differential diagnosis, while NN achieved the highest DDP of 99.6%, demonstrating it's proficiency in dealing imbalanced dataset. However, a lower DDR means it's less powerful in including all the pathology shown in the ground truth differential diagnosis. The modified Trans-

former has the higher metrics comparing with the original one. LSTM achieved the best performance in almost all metrics except DDP.

Table 3 shows the comparison of auto-diagnosis (AD) system performances between the two AD systems mentioned in the DDXPlus (Fansi Tchango et al., 2022), AARLC and BASD, which models were built based on (Yuan & Yu, 2024) and (Luo et al., 2021), and our auto-diagnosis system built by the LSTM model. The AD-LSTM model has the IL value equal to 20 without std value. The result of Table 3 indicates that overall, AD-LSTM has the best performances among the three when it comes to DDR, DDF1, and GM. While in the respect of DDP, AARLC still has the highest score.

7 Discussion

We changed the model structure in order to better predict and mimic the real world experience. In the previous papers, (Fansi Tchango et al., 2022) and (Alam et al., 2023), the Transformer model simply input all the variables without implementing the inquiry system, which make the model less applicable in real healthcare system. Although the AARLC and BASD implemented the inquiry system, their GM showed that these 2 models are not robust enough.

To deal with the problems above, we chose different methods to predict pathology and differential diagnosis. Then, we incorporated a simple inquiry system with the best-performance model of differential diagnosis prediction to build an auto-diagnosis system. The LSTM model outperformed the neural network models, which might be attributed to the capability to extract useful information from a long input. In addition, we also found that the LSTM model performed better than the transformer-based model. This is probably because the order of pathologies in differential diagnosis was considered in the transformer-based model, while the LSTM model simply treated them as multi-label classification. For our inquiry system, it actually performed well comparing to neural network based system. However, because of the entropy calculation of the dataset, it requires more space and time to process the data. During the experiment, there are some limitations. Considering the size of the datasets and the time consumption of each model, we only used grid search to choose the best combination of parameter in a small range. Also, because we do not have access to the AD system that generate the questions, we have to drop the `INITIAL_EVIDENCE` column to avoid any confusion. Another constraint is from our inquiry system. The question length in our inquiry system is predefined by setting it to be 20. However, the question length ideally should be short when the case is a general case and longer when it is a tricky case.

In future work, we plan to implement the AD system to more complex datasets, which can imitate real-world experience and better assist doctors while deciding the order of questions.

8 Conclusion

Our work used different models to predict pathology and differential diagnoses. We applied the best model of differential diagnosis prediction to a simple inquiry system to build an auto-diagnosis system. The key achievement is that we showed that a modified transformer model and an LSTM architecture could achieve higher metrics during differential diagnosis prediction, showing machine learning models' robustness in handling large imbalanced medical data. The AD system we built also demonstrated improvement in metrics to previous researches. Models using the technique of Random Forest demonstrated exceptional metrics, highlighting their potential in practical medical applications. Overall, our project showed that there is still much work that can be done by machine learning in the healthcare area and again proved the importance of it.

References

Alam, M. M., Raff, E., Oates, T., & Matuszek, C. (2023). Ddxt: Deep generative transformer models for differential diagnosis. *Deep Generative Models for Health Workshop NeurIPS 2023*.

Dong, Y., Cordonnier, J.-B., & Loukas, A. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *International Conference on Machine Learning*, 2793–2803.

Eftimie, L. G., Glogojeanu, R. R., Tejaswee, A., Gheorghita, P., Stanciu, S. G., Chirila, A., Stanciu, G. A., Paul, A., & Hristu, R. (2022). Differential diagnosis of thyroid nodule capsules using random forest guided selection of image features. *Scientific Reports*, *12*(1), 21636.

Fansi Tchango, A., Goel, R., Wen, Z., Martel, J., & Ghosn, J. (2022). Ddxplus: A new dataset for automatic medical diagnosis. *Advances in Neural Information Processing Systems*, *35*, 31306–31318.

Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. *New England Journal of Medicine*, *388*(13), 1201–1208.

Huang, C., Chen, W., Liu, B., Yu, R., Chen, X., Tang, F., Liu, J., & Lu, W. (2022). Transformer-based deep-learning algorithm for discriminating demyelinating diseases of the central nervous system with neuroimaging. *Frontiers in immunology*, *13*, 897959.

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2020). Ai in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, *3*, 4.

Luo, H., Li, S.-W., & Glass, J. (2021). Knowledge grounded conversational symptom detection with graph memory networks [arXiv preprint arXiv:2101.09773].

Men, L., Ilk, N., Tang, X., & Liu, Y. (2021). Multi-disease prediction using lstm recurrent neural networks. *Expert Systems with Applications*, *177*, 114905.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), 1236–1246.

Sunena Rose, M., & Sobhana, N. (2021). Automated diagnosis of diseases using integrated machine learning approaches. *International Conference on Soft Computing and Pattern Recognition*, 195–204.

Yuan, H., & Yu, S. (2024). Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *Artificial Intelligence in Medicine*, *148*, 102748.