MediConv: A Covid-19 Conversation Dataset with Entity Labels for Dialog Systems

Daksh Dadhania, Zaid Khan, Parva Chowdhury
Manipal Academy of Higher Education
Email: dakshdadhania@gmail.com, kzaidnba@gmail.com, parva.chowdhury@gmail.com

***Abstract*—The ongoing COVID-19 pandemic had necessitated the rapid development of remote healthcare solutions. In this context, intelligent dialog systems serve as a crucial interface for disseminating medical information and providing user-friendly consultation. This paper introduces "MediConv," a comprehensive dataset comprising over 7,000 annotated conversations that encompass a wide range of COVID-19-related medical scenarios. Each conversation is meticulously labeled with medical entities such as symptoms, diagnoses, and treatments to facilitate the training of conversational agents. We suggest a fresh method for improving dialog systems using entity-driven conversation modeling that aims to generate more accurate and contextually relevant responses. We evaluate various machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forests, and Multilayer Perceptrons (MLP),XgBoost,LightGBM with K fold cross validation,Ensemble Model of Random Forest and XgBoost to predict the next entity in a conversation sequence and generate appropriate responses. Our findings suggest that the Random Forest and MLP models, in particular, show superior performance on the F-measure, indicating their potential in developing effective dialog systems for medical applications. The "MediConv" dataset not only serves as a benchmark for future research in medical Natural Language Processing (NLP) but also lays the foundation for AI-powered healthcare systems capable of managing patient interactions during widespread health emergencies. In addition to the previously described methods, advanced machine learning techniques such as XGBoost, ensemble methods, and LightGBM were also evaluated to further enhance the dialog system's performance.**

***Index Terms*—COVID-19, dialog systems, machine learning, dataset, conversation modeling**

I. Introduction

The COVID-19 pandemic has unleashed unprecedented challenges on a global scale, stretching healthcare systems to their limits and necessitating innovative approaches to medical consultation and patient care. With the infectious nature of the virus, conventional face-to-face medical consultations pose risks to both healthcare providers and patients. Consequently, there has been a marked shift towards telehealth solutions and AI-driven dialog systems that offer remote consultation services. These systems not only reduce the risk of infection but also address the surge in demand for medical advice and diagnostics, thereby alleviating the burden on healthcare facilities.

Nevertheless, there are many obstacles in the way of using dialog systems in the medical field. Due to their inherent complexity and sensitivity, medical talks demand a high level of accuracy as well as context awareness. Moreover, these systems must be both resilient and flexible in response to new data due to the COVID-19 pandemic's quick evolution, which is marked by new symptoms and treatment guidelines. For this reason, specific datasets that represent the subtleties of COVID-19-related medical dialogs are crucial for both system evaluation and training.

"MediConv," introduced in this paper, represents a significant step towards this goal. The dataset has been carefully curated to include over 7,000 dialogues that are richly annotated with medical entities relevant to COVID-19. The construction of this dataset was guided by the pressing need for conversational agents that can handle a spectrum of queries, ranging from general information about the virus to specific medical advice tailored to the symptoms and history of the individual.

The research delineated in this document is structured around the following key objectives: First, to create a targeted dataset that can bridge the gap between general conversational models and the specialized requirements of medical dialog systems. Second, to develop a suite of predictive models that can anticipate the flow of conversation in a medical context and generate responses that are contextually and medically relevant. Third, to conduct a comprehensive evaluation of these models using both automatic metrics and human judgment to ensure that the generated responses meet the high standards required for medical advice. Lastly, the overarching aim is to contribute to the knowledge base of Natural Language Processing (NLP) and Artificial Intelligence (AI) in the healthcare domain, showcasing the applicability of advanced dialog generation techniques to real-world crises.

In developing "MediConv," we have addressed several gaps in the current landscape of medical NLP research. While there are existing datasets for dialog systems, few are dedicated to the intricacies of COVID-19-related conversations. Moreover, those that do exist often lack the depth of annotation necessary to train sophisticated models. "MediConv" fills this void by providing detailed annotations that include not only medical entities but also the intent and sentiment behind the conversations. This allows for the development of models that can understand and empathize with the user, a critical component in medical dialogs.

The methodological contributions of this paper are manifold. We adopt a multifaceted approach to model development, incorporating techniques from various areas of machine learning. The K-Nearest Neighbors (KNN) model provides a baseline for performance, offering insights into the value of similarity-based reasoning in conversation. The Support Vector Machine (SVM) introduces the power of kernel-based learning, capturing non-linear patterns within dialog flows. Decision Trees and Random Forests reveal the importance of feature selection and the benefits of ensemble learning. Finally, the Multilayer Perceptron (MLP) showcases the capabilities of neural networks in capturing the subtleties of human language.

Through rigorous evaluation, we demonstrate that while traditional models like KNN provide a solid foundation, it is the more sophisticated Random Forest and MLP models that truly excel in this domain. Their ability to factor in a multitude of variables and learn complex representations makes them particularly well-suited for medical dialog systems.

This paper is organized as follows: Section II reviews related work in the field of medical dialog systems and datasets. Section III details the creation of the "MediConv" dataset and the methodology employed in annotating the conversations. Section IV describes the various machine learning models we developed and the rationale behind their selection. Section V discusses the evaluation metrics used and presents the results of our experiments. Section VI analyzes these results and their implications for future research and application. Finally, Section VII concludes the paper with a summary of our findings and a look ahead to ongoing and future work in this exciting and vital area of research.

II. Related Work

A. Literature Survey

Paper[1] tackles a crucial challenge in the realm of medical conversational agents, particularly noted during the COVID-19 pandemic. The authors highlight the lack of extensive medical dialogue data, which hampers the development of effective end-to-end neural-based dialogue systems for healthcare. To address this shortfall, they introduce a significant medical dialogue dataset called MedDG, which includes over 7,000 conversations related to two common gastrointestinal diseases. This dataset is richly annotated with categories such as diseases, symptoms, attributes, tests, and medications, providing a robust foundation for further research. The paper proposes two specific tasks using the MedDG dataset: predicting the next entity in a dialogue and generating doctor responses. The authors test these tasks against current benchmarks and develop two specialized dialogue models that incorporate entity prediction. Their findings reveal that pre-trained language models and other baseline methods underperform, suggesting a substantial potential for improvement, particularly through the integration of auxiliary entity data. A simple retrieval model demonstrated superior performance over more complex generative models in human evaluations, underscoring the significant potential for further advancements in creating more accurate and clinically relevant responses. The authors advocate for further development of sophisticated prediction and generation methods, expansion of the dataset to include more disease types and scenarios, and the implementation of more stringent evaluation techniques to better assess the effectiveness and applicability of conversational agents in healthcare.

Paper [2] tackles a notable challenge in task-oriented dialog (TOD) systems, which frequently rely on external knowledge bases (KBs) to retrieve necessary information, like restaurant details, for generating responses. The prevailing methods either explicitly retrieve data from the KB or implicitly integrate it into the model's parameters. The explicit method, though accurate, becomes inefficient as the KB size increases, while the implicit method is more flexible and efficient. However, both can lead to inconsistent entity information in responses. To mitigate this, the authors suggest an autoregressive entity generation approach that uses this initial entity data to steer the rest of the response generation in an end-to-end system. They also employ a trie-based constraint for entity consistency and a logit concatenation strategy to improve gradient backpropagation during training. Their experiments on the MultiWOZ, SingleWOZ, and CAMREST datasets demonstrate improved quality and consistency in generated responses. They propose further investigation into advanced entity prediction techniques, broader datasets encompassing various medical scenarios, and enhanced evaluation methods to better determine the responses' effectiveness in medical contexts.

Paper [3] presents a medical dialogue system addressing the urgent need for Covid-19-related consultations during the pandemic. The literature survey highlights previous research in medical dialogue generation, transfer learning in NLP, Covid-19 dialogue datasets, dialogue generation models, and evaluation metrics for dialogue systems. Leveraging transfer learning techniques and CovidDialog datasets, the proposed system aims to generate doctor-like responses that are relevant and clinically meaningful. Evaluation metrics including automatic and human evaluation demonstrate the effectiveness of the approach in providing Covid-19-related consultations. The availability of data and code promotes reproducibility and further research in the field.

Paper [4] introduces a novel dialogue model, X-ReCoSa, addressing the limitation of conventional hierarchical dialogue models which typically ignore sentence-level representations from word-level encoders, focusing only on utterance-level context. This often results in information loss during response generation. The X-ReCoSa model is designed to integrate multi-scale contextual information, utilizing both sentence and utterance representations. The model comprises an upper "intention" part and a lower "generation" part. The former processes the context to set the response's intention, while the latter generates the actual words based on sentence-level input. This dual-layer approach effectively merges hierarchical context into response generation. The model's effectiveness was validated on the DailyDialog dataset, showing superior performance to baseline models through both automated and human evaluations.

Paper [5] highlights the issues practical dialog systems face, such as handling diverse knowledge sources, interpreting noisy user inputs, and the lack of sufficiently annotated data, especially for systems designed for the Chinese language. To fill this research gap, the authors introduce CGoDial, a comprehensive benchmark for evaluating Chinese goal-oriented dialog systems. This benchmark includes a variety of tasks and domains, offering a robust platform for testing. Additionally, they propose a new metric, Dialog Success F1 (DSF1), which evaluates both task completion and dialogue quality, aiming to provide a more precise assessment of system performance. Experimental results affirm the benchmark's utility and the effectiveness of the new evaluation metric, setting the stage for further advancements in Chinese dialog systems.

Paper [6] discusses enhancing dialogue generation models by incorporating fine-grained context and knowledge weighting during model training. Recognizing the importance of conversational context and external knowledge in crafting relevant responses, the authors develop a model that optimally balances these elements to avoid the inclusion of extraneous information. The model was trained using a large dialogue dataset and assessed with several metrics, which confirmed its efficacy in refining dialogue quality. This novel context and knowledge weighting method could significantly improve dialogue system performance.

Paper [7] explores improving knowledge-grounded dialogues (KGD) by refining how knowledge snippets are selected. Traditional methods, which often treat snippets as isolated relevant or irrelevant items, fail to recognize their collective interplay and the overarching discourse structure. To address these shortcomings, the authors propose GenKS, a generative sequence-to-sequence model that uses an attention mechanism to better integrate and utilize knowledge snippets in dialogues. This method also includes a hyperlink mechanism to explicitly model interactions between the dialogue and embedded knowledge. Tests on three benchmark datasets show that GenKS surpasses previous methods in knowledge selection and response generation, markedly enhancing dialogue quality and the overall effectiveness of dialogue systems.

Paper [8] introduces a novel approach to enhancing dialog generation systems by incorporating speaker history and context into the generation process. This method aims to produce more coherent and contextually relevant responses in dialog systems, addressing a common challenge in natural language processing (NLP) related to maintaining context and relevance over the course of a conversation. The approach leverages contrastive learning and prompt-based techniques to better understand and utilize the history of interactions with a specific speaker. By doing so, SHADE aims to improve the dialog system's ability to generate responses that are not only contextually appropriate but also personalized to the speaker's history and preferences. This represents a significant advancement in dialog generation technology, moving beyond generic responses to more tailored and engaging interactions. In the broader context of NLP and dialog systems research, SHADE's methodology aligns with ongoing efforts to enhance the naturalness and relevance of machine-generated text. The use of contrastive learning and prompt-based methods reflects a growing trend in the field towards more sophisticated models that can better mimic human-like conversational abilities. These techniques have shown promise in various NLP tasks, including few-shot learning, text classification, and information extraction, by enabling models to learn more effectively from limited data and to generate more accurate and context-aware outputs. The significance of SHADE lies in its potential to improve the user experience in applications involving conversational agents, such as virtual assistants, customer service bots, and interactive entertainment systems. By making dialog generation more context-aware and personalized, SHADE could lead to more engaging and satisfying interactions between humans and AI systems. In conclusion, "SHADE: Speaker-History-Aware Dialog Generation Through Contrastive and Prompt Learning" presents a promising approach to overcoming some of the longstanding challenges in dialog generation. Its focus on leveraging speaker history and context through advanced learning techniques offers a pathway to more natural and effective conversational agents, contributing to the ongoing evolution of NLP technologies.

Paper [9] explores the advancement of goal-oriented dialog systems using deep neural networks. By applying end-to-end learning, Rajendran aims to construct dialog systems capable of assisting users in various tasks, such as booking flights or making restaurant reservations, without the need for domain-specific handcrafting. Addressing key challenges in this approach, including handling a large number of named entities, learning with multiple valid next utterances, adapting to new user behaviors, and overcoming the need for large training datasets, Rajendran proposes novel methods to improve system performance and adaptability. His work significantly contributes to the field of conversational AI by demonstrating the feasibility and benefits of end-to-end learning for goal-oriented dialog systems, paving the way for more versatile and efficient conversational agents. Additionally, Rajendran's research context underscores various efforts to enhance dialog systems' efficiency and adaptability, such as advancements in machine learning tools for discriminating human versus AI-generated text and exploring emergent gauge fields. Ultimately, "On End-to-End Learning of Neural Goal-Oriented Dialog Systems" represents a significant advancement in conversational AI, offering innovative solutions to longstanding challenges and contributing to ongoing discussions on the ethical and practical implications of AI in everyday life.

Paper [10] explores deep learning techniques' application in developing neural dialog systems. This comprehensive work, structured around four pivotal articles, advances dialog systems by focusing on neural network-based approaches to enhance dialog generation and understanding. The key contributions include evaluating neural dialog architectures' efficacy in capturing conversation nuances, proposing methods to improve response generation quality in open-domain dialog systems by considering dialog attributes, introducing a cost-effective data collection method for task-oriented dialog systems, and presenting an embedding-free technique for word representation to reduce memory footprint and enhance robustness. Sankar's research provides insights into challenges and opportunities in neural dialog modeling, addressing issues like insensitivity to context perturbations and the need for improved response generation techniques, while offering practical solutions for data collection and word representation. This work significantly contributes to advancing conversational AI, shaping future research directions, and offering methodologies applicable in natural language processing and machine learning domains.

Paper [11] introduces a pioneering approach to generating medical dialogues, crucial for efficient diagnosis and treatment assistance in healthcare. The key contributions include the introduction of latent variables for patient states and physician actions within an end-to-end variational reasoning framework, overcoming challenges posed by limited labeled data in medical scenarios. By employing a variational Bayesian generative model and a two-stage collapsed inference method, the system efficiently approximates posterior distributions and minimizes bias during training. Additionally, the development of a physician policy network enhances the system's reasoning ability, resulting in more accurate and contextually relevant medical responses. This approach represents a significant advancement in medical dialogue generation, addressing data scarcity challenges and demonstrating superior performance compared to state-of-the-art baselines across multiple evaluation metrics. Overall, "Semi-Supervised Variational Reasoning for Medical Dialogue Generation" contributes a novel framework with substantial implications for enhancing dialogue systems in healthcare, paving the way for further research and application in conversational AI and healthcare informatics.

Paper [12] is crucial for various medical applications. While existing approaches for general knowledge graphs exist, they often overlook the unique challenges posed by the medical domain. The survey highlights key contributions in medical entity relation verification, machine reading comprehension (MRC), and knowledge graph construction. In medical entity relation verification, Hirschman et al. (2005) initiated the BioCreAtIvE challenge, catalyzing research in biomedical relation extraction. Wang et al. (2019) proposed a distant supervision method for medical entity relation extraction, while Jianget al. (2020) developed a framework for medical relation extraction from Chinese clinical texts. In biomedical MRC, Lee et al. (2019) introduced BioBERT, a domain-specific pre-trained language model, and Beltagy et al. (2019) proposed SciBERT, enhancing performance in biomedical NLP tasks. For knowledge graph construction, Toutanova et al. (2015) introduced the OpenIE framework, enabling automatic extraction of relational tuples from text. Liu et al. (2019) proposed MedGNN, a graph neural network-based framework for medical knowledge graph completion. Challenges include the high variability of medical terms and difficulty in evidence retrieval from complex medical documents. Recent advancements leveraging domain-specific pre-trained language models and distant supervision show promise, but challenges like handling medical terminology complexity persist. Continued research and innovation are needed to address these challenges and further enhance medical entity relation verification.

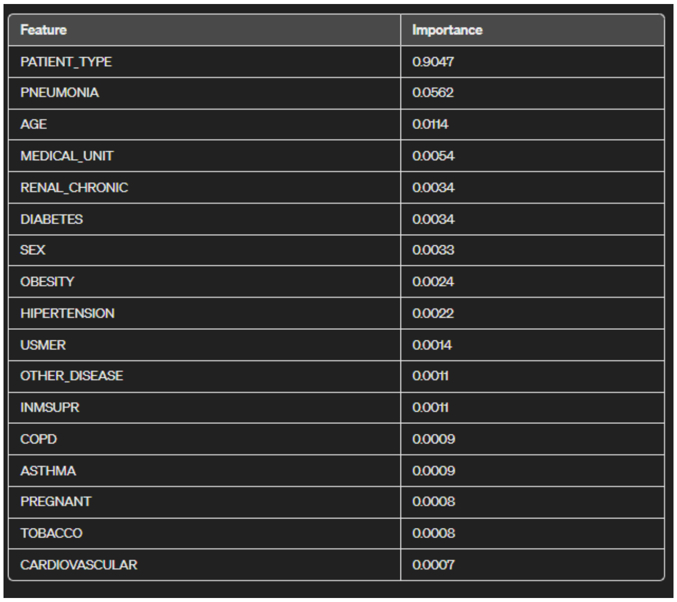
Paper [13] introduces a novel approach to conversational response generation, focusing on discourse-level coherence among utterances in dialogue contexts. This survey offers insights into related work in conversational response generation, pre-trained language models (PLMs), and discourse modeling. In conversational response generation, Vinyals et al. (2015) introduced Seq2Seq models with attention mechanisms, revolutionizing the field. Radford et al. (2019) proposed GPT for response generation, achieving state-of-the-art performance. Zhang et al. (2021) introduced UniLM, a unified PLM for response generation, leveraging bidirectional and unidirectional language modeling objectives. For PLMs tailored for dialogue, Devlin et al. (2019) introduced BERT, laying the foundation for subsequent research. Wolf et al. (2019) proposed DialoGPT, fine-tuned for dialogue generation, while Lewis et al. (2020) introduced BART for sequence-to-sequence learning tasks, including dialogue generation. In discourse modeling, Li et al. (2016) introduced a hierarchical RNN architecture for modeling discourse structure, improving coherence in dialogues. Zhou et al. (2018) proposed a discourse-aware neural response generation model, incorporating discourse-level features for generating contextually appropriate responses. Addressing limitations in existing PLM-based dialogue models, "DialogBERT" proposes a hierarchical Transformer architecture and training objectives tailored to recover and rank utterances based on discourse coherence. Outperforming existing baselines, DialogBERT demonstrates the effectiveness of incorporating discourse-awareness into conversational response generation, emphasizing the importance of coherence in dialogue systems.

Paper [14] delves into how the COVID-19 pandemic has reshaped the emotional evaluation of words, specifically examining whether it has altered the affective representation of COVID-19-related words compared to unrelated ones. Drawing upon a literature survey, the research underscores the crucial role of social context in shaping emotional processing and behavior, particularly highlighting the pandemic's significant impact on people's psychological and emotional well-being. It explores the interplay between social events, language, and emotional evaluation, emphasizing the scarcity of studies examining the pandemic's direct influence on word evaluation. The study employs experimental methodology to collect new ratings of valence and arousal for COVID-19-related and unrelated words, revealing noteworthy changes in the affective representation of these words during the pandemic. The findings underscore the dynamic nature of emotional representations and the profound influence of social context on language processing. Moreover, the study suggests future research avenues to explore the longitudinal evolution of affective representations of COVID-related words and individual differences in emotional processing, providing valuable insights into the pandemic's impact on the emotional evaluation of words.

Paper [15] a Covid consulting system is proposed to bridge the gap between patients and a limited number of doctors by integrating medical knowledge of Covid-19 with neural network generative models. A literature survey highlights the significance of pre-trained language models like BERT and GPT in healthcare-related NLP tasks, along with the surge in Covid-19 chatbots during the pandemic. Additionally, advancements in dialogue generation for healthcare and evaluation metrics for dialogue systems are discussed. The proposed system leverages fine-tuning of pre-trained models to automatically recognize Covid-19 symptoms and generate doctor-like responses. Evaluation metrics are employed to assess the system's performance, demonstrating its effectiveness in addressing the challenge of limited medical consultations during the pandemic.

III. Dataset Description

The primary objective of this project is to develop a machine learning model capable of predicting whether a Covid-19 patient, based on their symptoms, current health status, and medical background, is at high risk. To achieve this, we will employ various classification algorithms.
- K-Nearest Neighbours
- Support Vector Machine
- Decision Trees
- Multilayer Perceptron
- XGBoost
- Ensemble (Random Forest + XGBoost)
- LightBGM
- Random Forest



Feature	Importance
PATIENT_TYPE	0.9047
PNEUMONIA	0.0562
AGE	0.0114
MEDICAL_UNIT	0.0054
RENAL_CHRONIC	0.0034
DIABETES	0.0034
SEX	0.0033
OBESITY	0.0024
HIPERTENSION	0.0022
USMER	0.0014
OTHER_DISEASE	0.0011
INMSUPR	0.0011
COPD	0.0009
ASTHMA	0.0009
PREGNANT	0.0008
TOBACCO	0.0008
CARDIOVASCULAR	0.0007

Figure 1: Feature importance of Attributes

In each technique we will test a variety of hyperparameters values to get the best model. In addition, we would like to analyze how each feature, which will be detailed below, affects the chances of getting a severe illness from Covid-19, and consequently understand who the populations at increased risk are.
The dataset for this project obtained from Kaggle. It was provided by the Mexican government. This dataset contains a huge number of anonymized patient-related information including pre-conditions. The raw dataset consists of 40 different features and 1,048,576 unique patients. Since the description of the data and features names was in Spanish, We had to first translate all the feature names into English. Thereafter, the following actions were taken to make the data usable.

- All patients who haven't tested positive for COVID-19 were deleted.
- Features with unnecessary and irrelevant information have been deleted.
- For features which have many conclusive values all rows with inconclusive value were filtered.
- For features which have a very few conclusive values the entire feature deleted.
- All the data values modified to mainly ones and zeroes to get it converted into one hot vector.

IV. **Correlation Matrix**

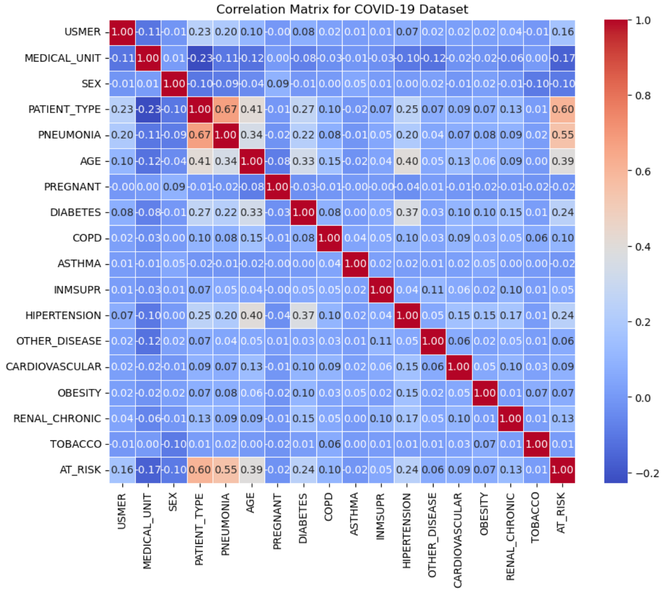


Figure 2: Correlation Matrix

After processing and cleaning, the dataset comprises 20 features (detailed below) and 388,878 unique patients. The entire data is divided into three groups: train (90%), validation (5%) and test (5%).

1. sex: female or male.
2. age: of the patient.
3. patient type: hospitalized or not hospitalized.
4. pneumonia: Indicates whether the patient already have air sacs inflammation or not.
5. pregnancy: Indicates whether the patient is pregnant or not.
6. diabetes: Indicates whether the patient has diabetes or not.
7. copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
8. asthma: Indicates whether the patient has asthma or not.
9. inmsupr: Indicates whether the patient is immunosuppressed or not.
10. hypertension: Indicates whether the patient has hypertension or not.
11. cardiovascular: Indicates whether the patient has heart or blood vessels related disease.
12. renal chronic: Indicates whether the patient has chronic renal disease or not.
13. other disease: Indicates whether the patient has other disease or not.
14. obesity: Indicates whether the patient is obese or not.
15. tobacco: Indicates whether the patient is a tobacco user.
16. usmr: Indicates whether the patient treated medical units of the first, second or third level.
17. medical unit: type of institution of the National Health System that provided the care.
18. intubed: Indicates whether the patient was connected to the ventilator.
19. icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
20. death: indicates whether the patient died or recovered.

The last three features serve as the label. That is, if a patient is intubed or treated in an intensive care unit or dies, he will be classified as at high risk (label 1).

Imbalanced Data

One of the significant challenges we will have to deal with is that the data is unbalanced. Only 15.4% (59,979) of the patients in the clean data are considered at risk while the rest of the patients are not at risk. As a result, we can not measure the quality of a model by its accuracy on the validation set, since it can always predict a patient is not at risk and his accuracy level will be 85%. Instead, we measure the quality of a model by F-measure, also called F1-score, which measures the accuracy of binary classification. The F-measure is the harmonic mean of the recall and precision.

Moreover, to avoid the algorithms we will use to develop a model that always predicts the patient is not at risk, there are several common methods we can use to solve this problem.
- Under-sampling - randomly delete examples from the majority class until the data is balanced.
- Over-sampling - randomly duplicate examples from the minority class until the data is balanced.
- Loss function intervention - Multiply in the loss function the percentage of error on the minority class by constant C which is the ratio between the number of objects in the data and the number of objects belonging to the minority class.

We use these methods in each of the machine-learning techniques and select the model with the best result.

The target class is AT_RISK which we predict. Its training data is generating by providing its label on the basis of ICU,DATA DIED and INTUBE attribute. In implementing data balancing strategies, it’s crucial to ensure that they are applied appropriately based on the nature of the data and the specific requirements of the predictive task. Each technique has its merits and pitfalls, and often the best approach is determined through experimentation and validation against a hold-out set or via cross-validation techniques.

The choice of strategy might also depend on the predictive model being used. For instance, tree-based models might handle imbalanced data differently from statistical models or neural networks. Thus, understanding the dynamics of each model with respect to imbalanced datasets is vital.

Finally, it's essential to monitor not just the F-measure but also other metrics like the The region beneath the receiver's operating characteristic curve (AUROC) and the Precision-Recall Curve. These metrics provide further insights into the trade-offs between different types of errors (false positives vs. false negatives) and the overall reliability of the model across various thresholds of classification.

By carefully managing the dataset's imbalance and selecting appropriate evaluation metrics, we can develop a model that not only predicts with high accuracy but also offers clinically relevant insights, ultimately enhancing decision-making in healthcare settings. Boosting algorithms such as AdaBoost can effectively increase the accuracy of combined weak classifiers and have been found to perform well with imbalanced datasets. Boosting focuses on sequentially correcting the mistakes of weak classifiers by focusing more on difficult-to-classify instances and less on those already handled well, naturally adapting to the imbalance in the dataset. Random Forests, a type of bagging technique, involves constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. It is effective because it reduces variance and helps avoid overfitting, which is particularly useful in the context of imbalanced datasets. When dealing with imbalanced datasets, using simple random splits for cross-validation can result in training folds that do not adequately represent the minority class. Stratified cross-validation ensures that each fold reflects the overall distribution of each class, leading to more reliable and robust model evaluation.

Methodology

A. K-Nearest Neighbors

The K-Nearest Neighbours is non-linear classifier in which an object is classified to the class most common among its k nearest neighbours in the train set. The hyperparameters for the KNN algorithm are the value of k and the distance function that calculates the distance between two objects. We implemented the KNN algorithm and used this implementation here as well. The prediction process in KNN is very slow so to speed it up I decided to use a distance matrix which is a matrix of shape [num validation, num train] where cell [i, j] hold the distance between the i-th validation point and the j-th training point. The main difficulty was how to calculate the distance matrix by each distance function. Fortunately, we found a function in the “scipy.spatial” library that calculates a distance matrix and gets a p value for the distance function as a parameter. Another difficulty was how to implement the algorithm as a generic class that could be run on any dataset. After many attempts I did it successfully.

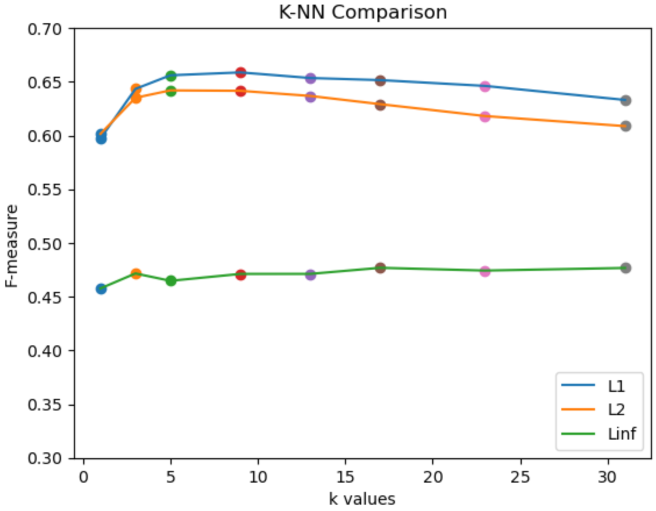


Figure 3: KNN Classifier Comparison

The k values we chose to check are k=[1, 3, 5, 9, 13, 17, 23, 31]. Each of this k values are tested on three distance functions: Manhattan distance (L1), Euclidean distance (L2) and Fréchet distance (L). Since KNN is very slow, the training set can not be larger than a few thousand. Therefore, the only method we can use to balance the data is undersampling. The results of all these different models on the validation set are shown in the following figure. The best model uses the L1 distance function with k = 9. This model yielded F-measure = 0.64.

L yielded the worst results. The reason for this is that the distance in L is calculated according to only one feature while our data includes many features that are all worth considering simultaneously. In addition, as the value k increases the F-measure slightly decreases. This can be explained by the fact that the data includes many points from both classes that are close to each other, so the more neighbors we consider the more likely we are to find neighbors who are not from the patient class. As expected, the KNN model results are not good enough, as our data includes many patients with the same symptoms but at a different level of risk.

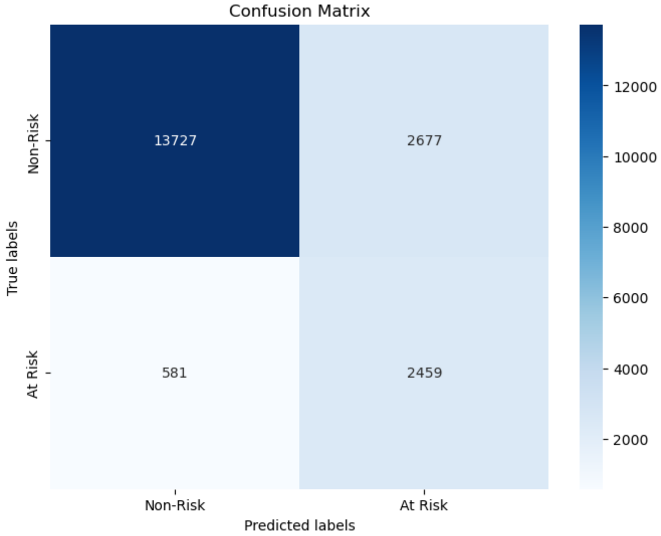


Figure 4: KNN Confusion Matrix

B. *Support Vector Machine*

SVM is a classifier that separates data points using a linear hyperplane with the largest amount of margin between the two classes in the train set. Support Vector Machines (SVMs) are adept at executing non-linear classifications using the kernel trick. This method allows for the implicit transformation of input data into higher-dimensional feature spaces, facilitating effective classification.

I used the “sklearn.svm” library to run the SVM algorithm on the data. I chose to compare between all the kernels that this library allows, which are: linear, polynomial, and RBF (Radial Basis Function) kernels. Since SVM is very slow on large training set we can not use oversampling to balance the data. Therefore, I ran all the models once with undersampling the data and again with loss intervention built-in “sklearn.svm”.

The results of all these different models on the validation set are shown in the following figure.

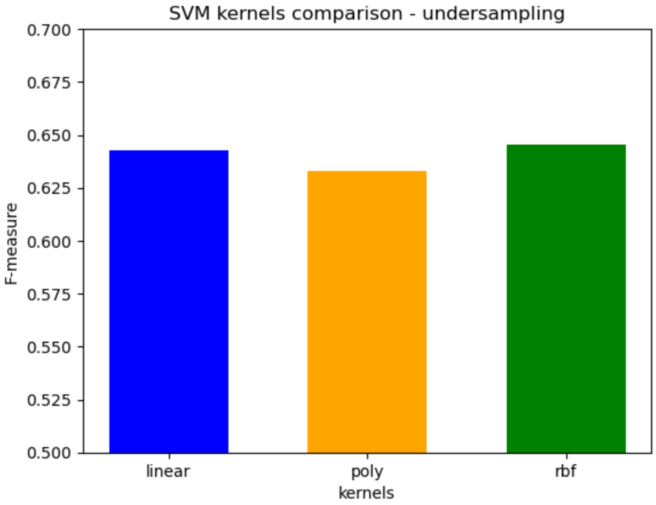


Figure 5: SVM Kernels comparision

The best model uses the RBF kernel with undersampling the data. This model yielded F-measure = 0.65.

Surprisingly, there is no significant difference between the two data balancing methods nor between a linear classification and a non-linear classification. A possible reason for this is since most of the data is represented as zero’s and one’s linear planes are sufficient to separate the two classes.

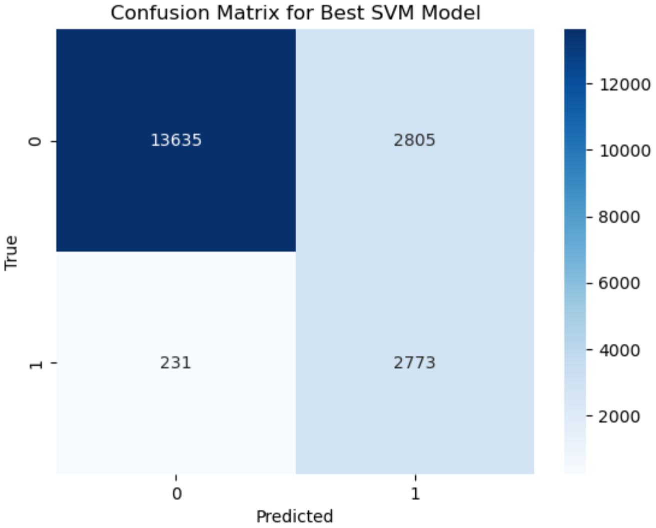


Figure 6: SVM Confusion Matrix

C. *Decision Trees*

A decision tree functions as a flowchart, where each internal node signifies a feature test, each branch the result of that test, and each leaf node represents a class label. Decision tree algorithms improve data splits at each node based on heuristic functions, which aim to optimize the separation of data for clearer classification paths

and rules.
we used the “sklearn.tree” library to run the decision tree algorithm on the data. I chose to compare between the two heuristic functions that this library allows, which are: ID3 and Gini index. For each of these heuristics you can choose the best split or the best random split. In addition, I chose to compare several values of the maximum tree depth that we allow the algorithm to build. The values tested are [5, 7, 11, 13, 17] . The maximum possible depth is 17 as this is the number of features we have in the clean dataset. The results of all these different models on the validation set are shown in the following figure. The best model uses the random id3 heuristic functions with maximum depth of 13 and loss intervention to balance the data.

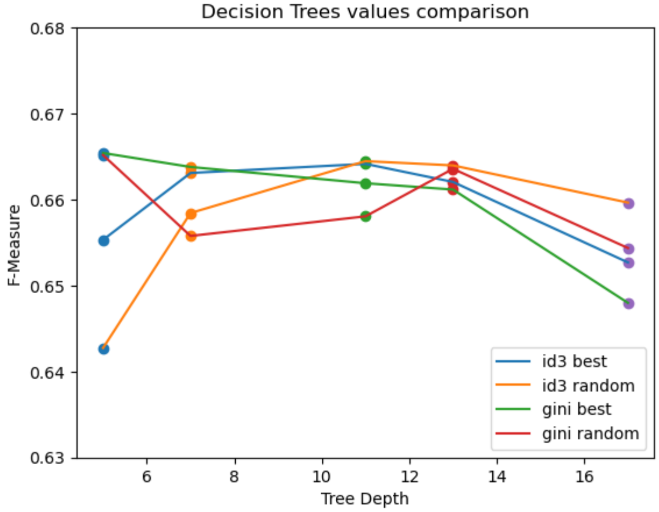


Figure 7: Decision Tree Comparision

Up to a certain depth the results improve but from this depth the results start to drop. This is because the tree needs to be deep enough to be able to separate all the objects from the two classes. Although from this depth onwards there is a decrease because the VC dimension in decision trees also depends on the depth of the tree. According to the generalization bound a large VC dimension reduces the quality of the model.
By plotting the tree of the best model, we can conclude that the most important features that affect a patient’s risk from covid-19 are his age and whether he suffers from pneumonia which mean air sacs inflammation.
We took patient dialogue and converted this into parameterized attributes for the model to predict.This was done through NLP and other algorithms.

D. *Random Forest*

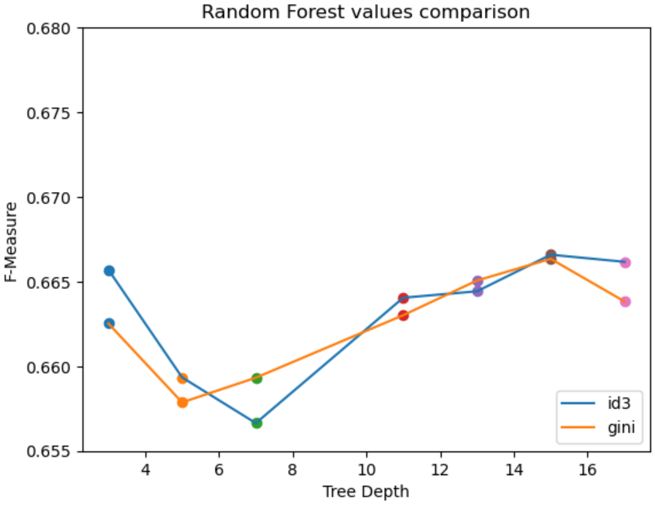


Figure 8: Random Forest Values
We also tried a model of a random forest using the “sklearn.ensemble” library. I tested this algorithm with the heuristics id3 and Gini index for the same depths mentioned above in decision trees. The results of all the models are shown in the following figure. The best model in decision trees is also the best model in random forest and as expected even improves it.

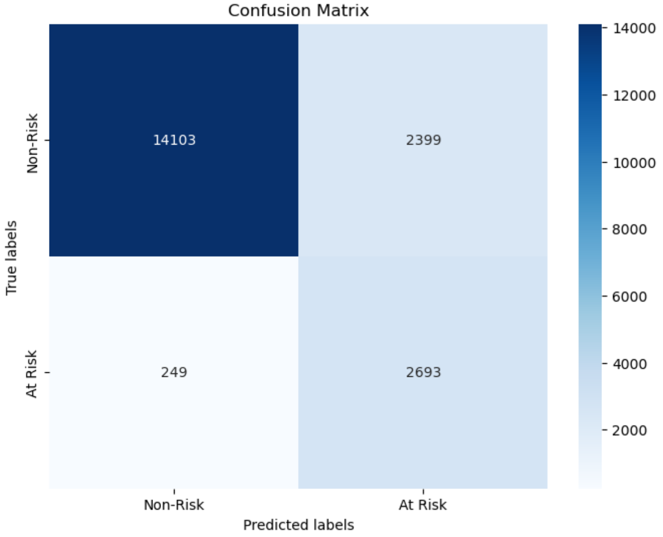


Figure 9: Random Forest Confusion

E. *Multilayer Perceptron*

A Multi-Layer Perceptron (MLP) comprises multiple layers, including an input layer, one or more hidden layers, and an output layer. Each node, or neuron, in the hidden and output layers employs a non-linear activation function, enhancing the model’s ability to learn complex patterns. The MLP model is trained using a supervised learning

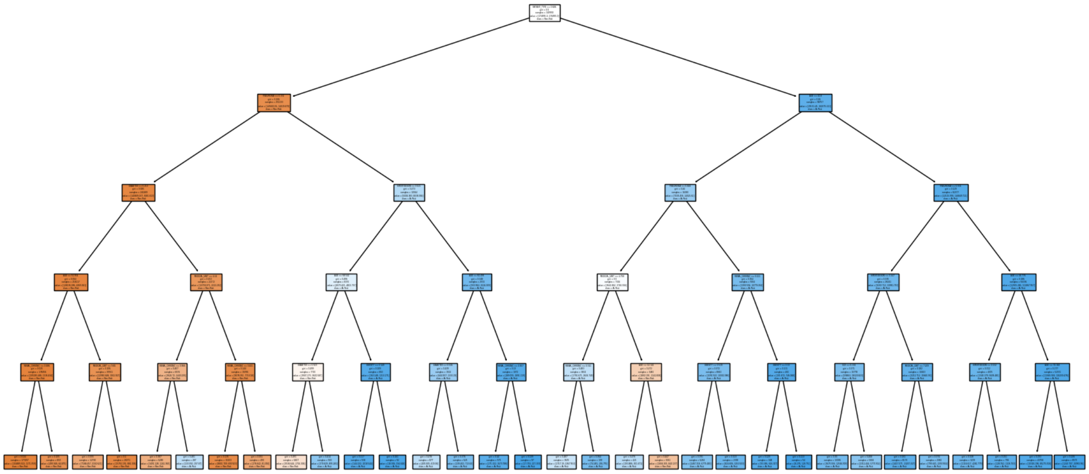


Fig. 1. 1: Decision Tree

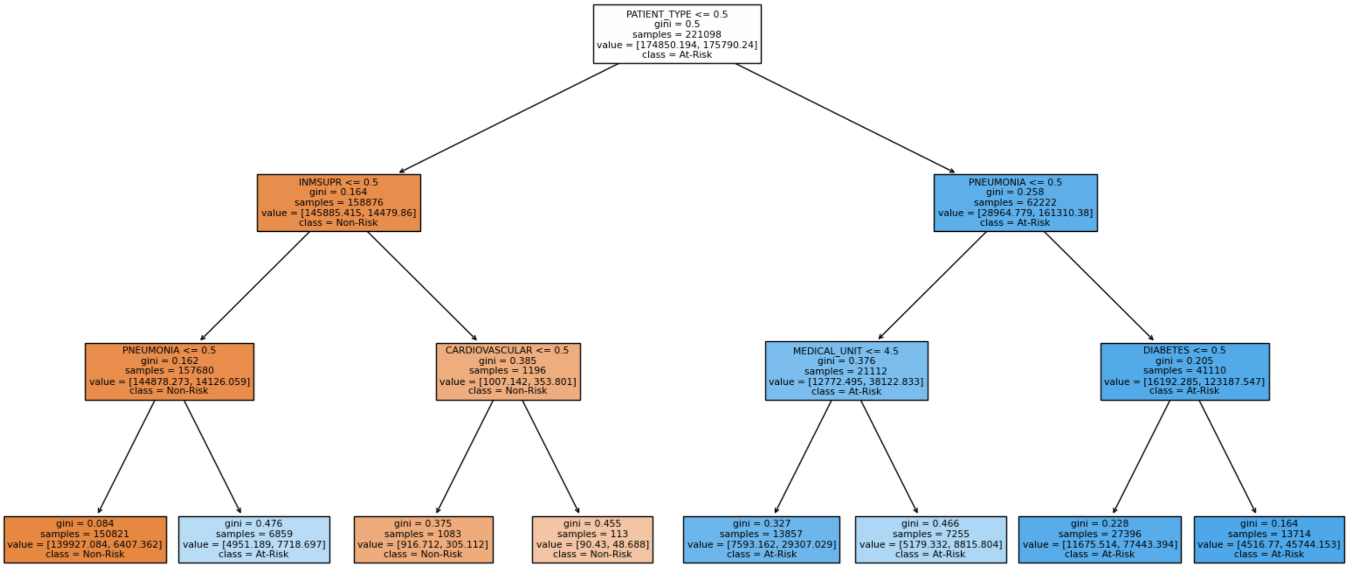


Fig. 2. 1: Random Forest Tree

technique known as backpropagation, which adjusts the model's weights based on the error rate of outputs compared to the true data labels.

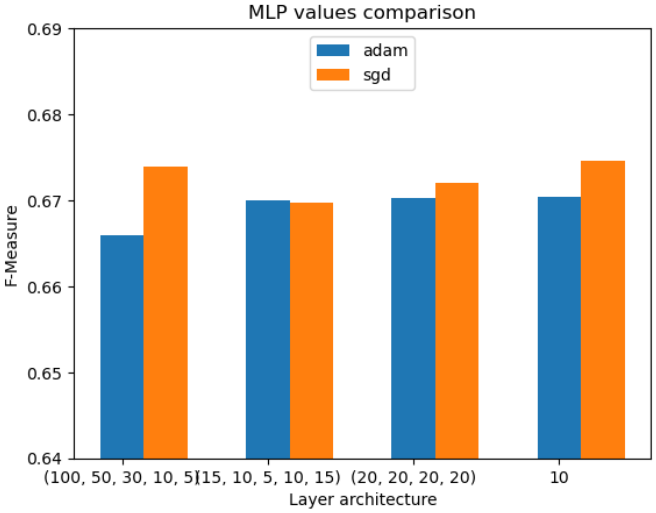


Figure 10: MLP Values Comparision

I used the “sklearn.neuralnetwork” library to run the MLP algorithm on the data. I chose to compare several layers of architectures on the two optimizers, Adam and SGD. The layers architectures tested are [(100,50,30,10,5), (15,10,5,10,15), (20,20,20,20), (10)]. Since the MLP algorithm works very slowly on large data, we cannot balance the data by using oversampling, so I used undersampling. The results of all these different models on the validation set are shown in the following figure. The best model uses Adam optimizer and only one hidden layer with 10 neurons.

Interestingly, the simplest neural network produced, with a big difference, the best results. This outcome shows that there are no deep connections between the various features in the dataset. That is, a combination of several features, which represent in the data different diseases, do not necessarily affect a patient’s being at risk from Covid-19. In addition, the MLP model provides good results. This outcome reinforces our conclusion from the results of the SVM models, that a planar classifier is good enough on our data.

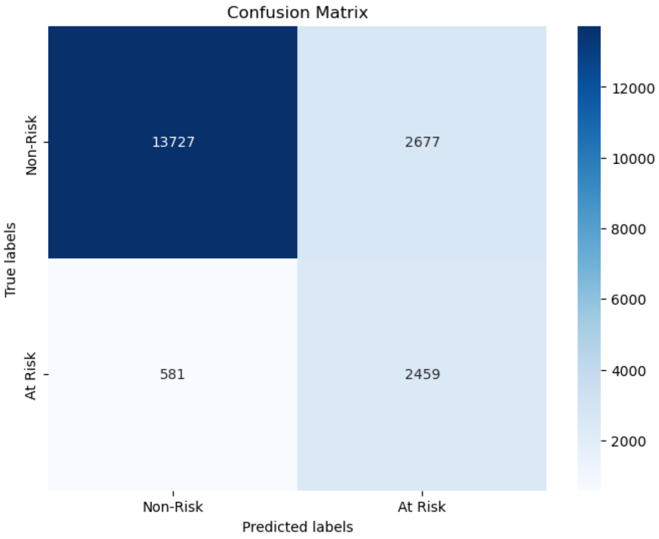


Figure 11: MLP Confusion Matrix

F. *XGBoost*

XGBoost, which stands for Extreme Gradient Boosting, is a scalable and accurate implementation of gradient boosting machines. Renowned for its performance and speed, XGBoost has become a widely used algorithm in machine learning competitions and practical applications. The core of the XGBoost algorithm lies in its ability to perform parallel tree boosting, which makes it highly efficient for large datasets.

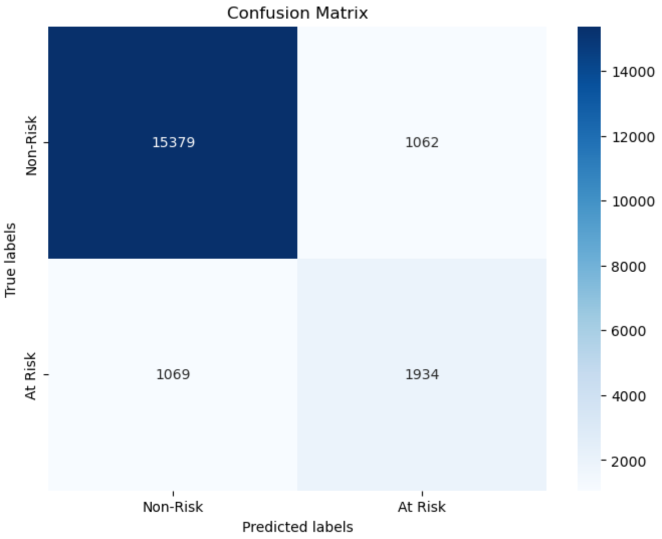


Figure 12: XGBoost Confusion Matrix

In our study, XGBoost is employed to tackle the classification of COVID-19 patient risk levels based on a myriad of clinical features. The gradient boosting framework of XGBoost allows it to iteratively refine the models by focusing on the difficult to classify instances, thus leading to a robust predictive performance. Its ability to handle missing data and its built-in regularization prevent overfitting, making it a suitable choice for our dataset.

G. *Ensemble Methods*

Ensemble methods combine predictions from multiple machine learning algorithms to achieve better performance than could be obtained from any single model on its own. In our work, we utilized a stacking ensemble approach that leverages the strengths of diverse models. By training a meta-model on the predictions of base classifiers, we enhance the system’s predictive accuracy.

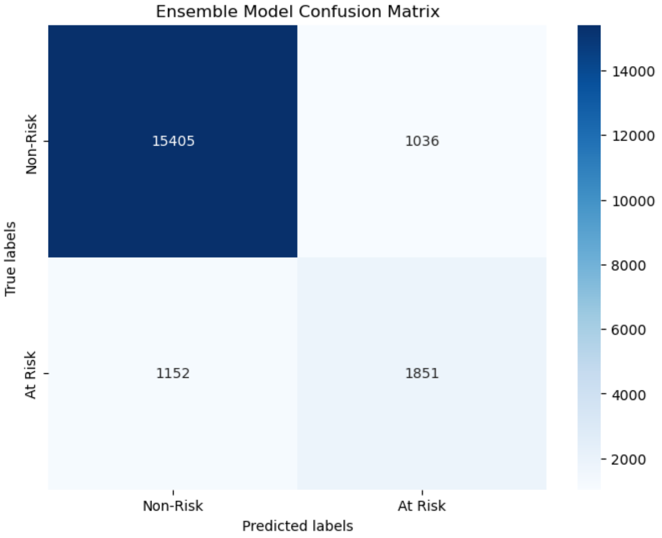


Figure 13: Ensemble Confusion Matrix

Specifically, our ensemble consists of a RandomForestClassifier and XGBoost, which were selected based on their individual performance on the validation set. A Logistic Regression model serves as the meta-learner, providing a final prediction based on the probabilities output by the base models. This approach capitalizes on the different patterns recognized by each base model, leading to a more accurate and stable prediction on the test set.

H. ***LightGBM***

LightGBM, which stands for Light Gradient Boosting Machine, is another gradient boosting framework that differentiates itself with its efficiency and scalability, especially on large datasets. Unlike other boosting frameworks, LightGBM grows trees leaf-wise (best-first) rather than level-wise and thus produces much deeper trees. Additionally, LightGBM implements an advanced form of histogram-based learning which speeds up training and reduces memory usage.

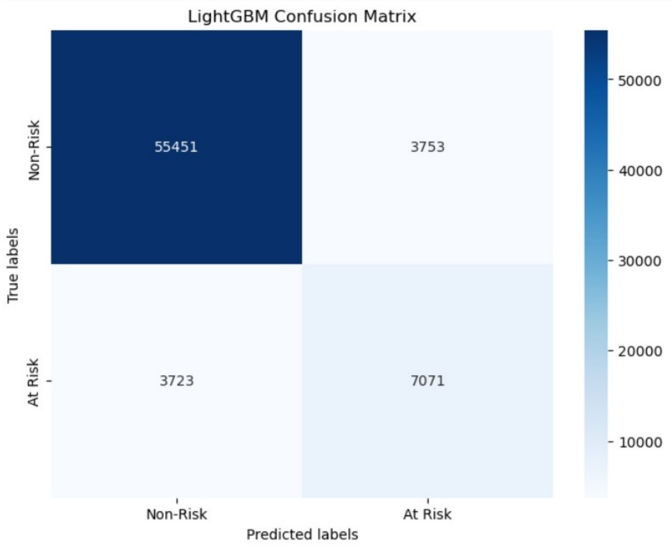
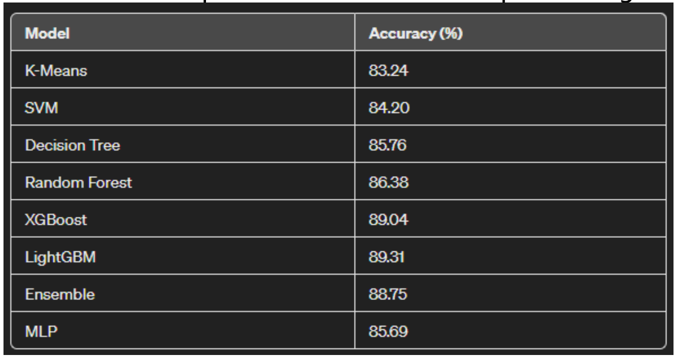


Figure-14 LightGBM Confusion Matrix

For our COVID-19 dataset, which is substantial in size, LightGBM is an ideal candidate due to its high efficiency and lower memory consumption. The framework’s ability to deal with categorical features and support for parallel and GPU learning made it an appropriate choice for our predictive tasks. We applied a 5-fold stratified cross-validation strategy within LightGBM to ensure that our model is robust and generalizes well to unseen data.

V. **Conclusion**

In the culmination of our study, we conducted a comparative analysis of model accuracies to assess the predictive capabilities of the models developed. Figure illustrates the accuracies of each model in a bar chart, where the x-axis represents the different models, and the y-axis represents their respective accuracies in percentages.



Model	Accuracy (%)
K-Means	83.24
SVM	84.20
Decision Tree	85.76
Random Forest	86.38
XGBoost	89.04
LightGBM	89.31
Ensemble	88.75
MLP	85.69

Figure-15 Model Accuracies

As depicted in the figure, LightGBM achieves the highest accuracy at 89.31%, closely followed by XGBoost with an accuracy of 89.04%. The ensemble model, which integrates the predictions of the Random Forest and XGBoost models using a meta-model, exhibits a strong performance with an accuracy of 88.747%. Random Forest independently achieves an accuracy of 86.381%, indicating its robustness as a standalone model.

Both Decision Trees and SVM perform commendably with accuracies of 85.76% and 84.2%, respectively. K-Means clustering, while not traditionally used for classification tasks, serves as a nonparametric baseline with an accuracy of 83.2%. The performance of each model underscores the effectiveness of ensemble and boosting methods in handling the complexity and intricacies of medical data related to COVID-19 patient risk classification.MLP gave 85.6accuracy.

In a performance comparison of machine learning models, Random Forest led with an F-measure of 0.67, followed by MLP at 0.664 and Decision Tree at 0.66. LightGBM and XGBoost scored 0.65 and 0.64 respectively, matching SVM’s accuracy of 0.64. The Ensemble model recorded 0.62, while K-Means had the lowest accuracy at 0.60.

These results demonstrate the value of advanced machine learning techniques in developing accurate and reliable predictive models for critical healthcare applications. The high accuracies achieved by the gradient boosting models, XGBoost and LightGBM, are particularly notable and validate their applicability in scenarios where precision is paramount. Also it was noticed that Random Forest gave the best F1 score.

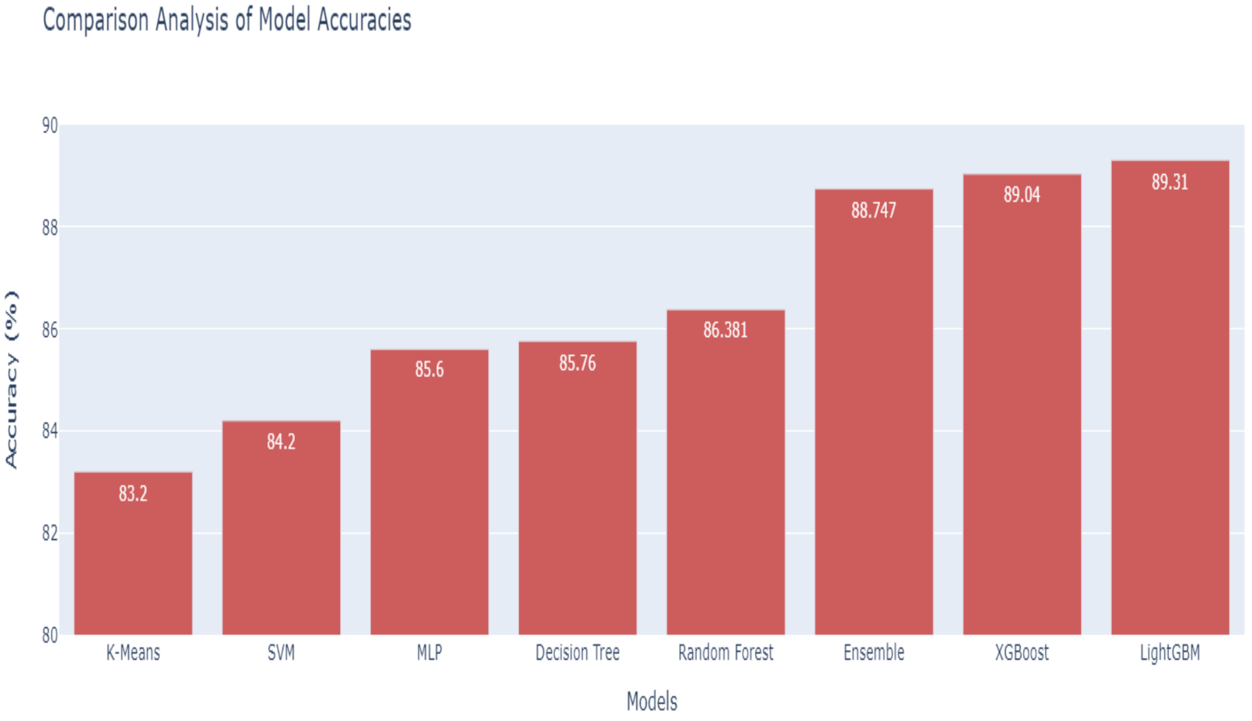


Figure-16 Comparison Analysis of Model Accuracies

VI. **Future Scope**

The advancements presented in this paper lay a foundational framework for further exploration and enhancement of dialog systems for medical applications, particularly in the context of the COVID-19 pandemic. The potential areas for future research and development are manifold and include both technological advancements and broader implementations. Here are some key areas of future scope: Integration of Multimodal Data:Future iterations of medical dialog systems could incorporate multimodal data such as images, voice inputs, and real-time physiological data from wearable devices. This integration would enable a more holistic approach to patient assessment and consultation, providing richer contexts for decision-making and advice.Expansion of Dataset Diversity:While "MediConv" provides a robust starting point, extending the dataset to include dialogues pertaining to other infectious diseases and medical conditions would enhance the system's applicability and resilience to future health crises. Additionally, including more languages and dialects could help in providing support to a broader user base.Advanced Natural Language Understanding (NLU):Leveraging advancements in NLU could improve the system's ability to interpret and respond to complex medical inquiries. Techniques like deep learning and transfer learning could be employed to better understand nuances and variations in patient communication.Personalized Healthcare Recommendations:Personalization algorithms could be refined to consider individual patient histories and demographic data to tailor advice and recommendations more accurately, thereby increasing the efficacy and safety of the advice provided.Real-Time Learning and Adaptation:Implementing real-time learning algorithms would allow the system to continuously learn from new dialogues and adjust its models accordingly. This adaptability is crucial in handling the evolving nature of diseases and medical knowledge.

These future directions not only promise to enhance the capabilities of medical dialog systems but also ensure their relevance and sustainability in an ever-changing healthcare landscape. The ongoing development in AI and machine learning, coupled with an acute understanding of medical needs, will continue to drive innovations that enhance patient care and healthcare delivery systems worldwide.

References

[1] W. Liu, J. Tang, Y. Cheng, W. Li, Y. Zheng, X. Liang, “MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation,” 2022.

[2] G. Huang, X. Quan, Q. Wang, “Autoregressive Entity Generation for End-to-End Task-Oriented Dialog,” 2022.

[3] W. Yang, G. Zeng, B. Tan, Z. Ju, S. Chakravorty, X. He, S. Chen, X. Yang, Q. Wu, Z. Yu, E. Xing, P. Xie, “On the Generation of Medical Dialogues for COVID-19,” 2022.

[4] D. Wu, “X-ReCoSa: Multi-Scale Context Aggregation For Multi-Turn Dialogue Generation,” 2023.

[5] Y. Dai, W. He, B. Li, Y. Wu, Z. Cao, Z. An, J. Sun, Y. Li, “CGoDial: A Large-Scale Benchmark for Chinese Goal-oriented Dialog Evaluation,” 2022.

[6] W. Zheng, N. Milic-Frayling, K. Zhou, “Contextual Knowledge Learning For Dialogue Generation,” 2023.

[7] W. Sun, P. Ren, Z. Ren, “Generative Knowledge Selection for Knowledge-Grounded Dialogues,” 2023.

[8] F. Wang, X. Zhao, X. Sun, “SHADE: Speaker-History-Aware Dialog Generation Through Contrastive and Prompt Learning.”

[9] J. Rajendran, “On End-to-End Learning of Neural Goal-Oriented Dialog Systems.”

[10] S. Chinnadhurai, “Neural Approaches to Dialog Modeling.”

[11] D. Li, Z. Ren, P. Ren, Z. Chen, M. Fan, J. Ma, M. de Rijke, “Semi-Supervised Variational Reasoning for Medical Dialogue Generation.”

[12] Y. Xia, C. Wang, Z. Shi, J. Zhou, C. Lu, H. Huang, “Medical Entity Relation Verification with Large-scale Machine Reading Comprehension.”

[13] X. Gu, K. M. Yoo, J.-W. Ha, “DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances.”

[14] C. Planchuelo, A. Baciero, “Social context effects on emotional language.”

[15] S. Tripathy, R. Singh, M. Ray, “Natural Language Processing for Covid-19 Consulting System.”