ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

NLP Based Prediction of Hospital Readmission using ClinicalBERT and Clinician Notes

LMatondora[1]; MMutandavari[1]; BMupini[2]
Information Security and Assurance Department[1], Computer Science Department[2]
School of Information Science and Technology, Harare Institute of Technology, Harare, Zimbabwe[1,2]
https://orcid.org/0009-0000-0830-0081[1] , https://orcid.org/0000-0003-3342-5710[1]

Abstract:- Hospital readmissions introduce a significant challenge in healthcare, leading to increased costs, reduced patient outcomes, and strained healthcare systems. Accurately predicting the risk of hospital readmission is crucial for implementing targeted interventions and improving patient care. This study investigates the use of natural language processing (NLP) techniques, specifically the ClinicalBERT model, to predict the risk of hospital readmission using the first 3-5 days of clinical notes, excluding discharge notes. We compare the performance of ClinicalBERT to other machine learning models, including logistic regression, random forest, and XGBoost, to identify the most effective approach for this task. This study highlights the potential of leveraging deep learning-based NLP models in the clinical domain to improve patient care and reduce the burden of hospital readmissions, even when utilizing only the initial clinical notes from a patient's hospitalization. It can also provide information early to allow Clinicians to intervene in patients who are at high risk. The results demonstrate that the ClinicalBERT model outperforms the other techniques, achieving higher accuracy, F1-score, and area under the receiver operating characteristic (ROC) curve. This study highlights the potential of leveraging deep learning-based NLP models in the clinical domain to improve patient care and reduce the burden of hospital readmissions.

Keywords:- *Hospital Readmission, Clinical Notes, ClinicalBERT, Deep learning,*

I. INTRODUCTION

Hospital readmissions pose a significant challenge to healthcare systems worldwide.[1] These unplanned returns within a short period of discharge (typically 30 days) [2]strain resources, increase costs, and negatively impact patient outcomes. Studies like [3]highlight the financial burden, estimating readmissions account for a substantial portion of healthcare spending. Research by [4]further emphasizes the negative consequences, suggesting a link between readmissions and increased mortality rates, as well as a decline in functional recovery for patients. These revolving-door readmissions not only affect individual patients but also create a ripple effect, impacting wait times and access to care for others[5].

Given these significant consequences, accurately predicting hospital readmissions within a specific timeframe is crucial[6]. Early identification of at-risk patients allows for the implementation of targeted interventions that can improve patient care, reduce healthcare costs, and optimize resource allocation[7]. Traditionally, readmission prediction models have relied on readily available structured data in electronic health records (EHRs) such as demographics, diagnoses, and past treatment history. While valuable, these models often overlook the wealth of information embedded within clinical notes.[8]

This study investigates the potential of Natural Language Processing (NLP) techniques, specifically the ClinicalBERT model, to predict hospital readmission risk using the first 3-5 days of clinical notes, excluding discharge summaries. ClinicalBERT is a pre-trained language model fine-tuned on a massive corpus of clinical text, enabling it to capture the nuances and domain-specific knowledge relevant to the healthcare domain.[9]

II. BACKGROUND TO THE PROBLEM/STUDY

Hospital readmissions have been a well-studied topic in healthcare research,[10] with numerous approaches proposed to address this challenge. Traditional machine learning models like logistic regression, random forest, and XGBoost have been commonly used for readmission prediction, utilizing structured patient data as input features [11], [12]These models offer valuable insights, but they often overlook the rich information within clinical notes.

Recent studies have explored the incorporation of unstructured data, such as clinical notes, to enhance the predictive performance of readmission models. For instance, [13]developed a deep learning model that combined structured and unstructured data to predict hospital readmissions, demonstrating the value of leveraging textual information for improved prediction accuracy.

Clinical notes, authored by healthcare professionals throughout a patient's hospitalization, offer a vast amount of textual data detailing the patient's medical history, current condition, clinical course, and social circumstances[14]. These narratives capture nuances often missed by structured data alone, such as subtle changes in a patient's condition, medication adherence concerns, or social determinants of health that could impact their recovery trajectory. By

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

leveraging NLP techniques, we can unlock the hidden insights within these clinical notes and potentially enhance the accuracy of readmission prediction[15], [16].

Recent studies have explored the incorporation of unstructured data, such as clinical notes, to enhance the predictive performance. For example[13] developed a deep learning model that combined structured and unstructured data to predict hospital readmissions, demonstrating the value of leveraging textual information.[17]

III. RELATED WORK

Hospital readmission prediction has been a well-studied area, with a recent shift towards leveraging Natural Language Processing (NLP) for improved accuracy. Traditional models relied on structured data like demographics, diagnoses, and treatment history, but these lack the rich context captured in clinical notes [11], [12].

Clinical notes offer valuable insights into a patient's medical journey, including social determinants of health, medication adherence concerns, and subtle changes in condition. Recognizing this potential, recent research has explored incorporating unstructured data like clinical notes into prediction models[18], [19]. Pioneering work by [13]demonstrated the effectiveness of combining structured and unstructured data. Their deep learning model achieved promising results in predicting readmissions, showcasing the value of NLP in this domain.

Further research has delved deeper into NLP techniques for analyzing clinical notes. Studies explored methods like word embeddings and recurrent neural networks (RNNs) to extract risk factors associated with readmission [20]–[22].The rise of pre-trained language models like BERT has opened new avenues for clinical NLP tasks. BERT's ability to understand complex language relationships has proven effective in tasks like diagnosis prediction and medication extraction [23], [24]. Building upon this foundation, ClinicalBERT, a BERT model specifically fine-tuned on a massive corpus of clinical text, has emerged as a powerful tool for various healthcare applications.

Huang [23]investigated the use of ClinicalBERT for predicting hospital readmissions, highlighting its potential for improved accuracy. Alsentzer et al. [24]further solidified this by demonstrating the effectiveness of ClinicalBERT embedding in various clinical NLP tasks. Their work provides a strong foundation for utilizing ClinicalBERT in hospital readmission prediction, which is the focus of this study. While traditional methods focused on structured data, the field has shifted towards incorporating the richness of clinical notes through NLP techniques[25]–[27]. Our study builds upon this evolving landscape by leveraging the advanced capabilities of the ClinicalBERT model to extract meaningful insights from clinical notes and enhance hospital readmission prediction accuracy.

IV. METHODOLOGY

This section details the data selection, model selection and training, and evaluation metrics employed in the study to predict hospital readmission risk using clinical notes and the ClinicalBERT model. The model was built using the ‘ADMISSIONS’ and ‘NOTEEVENTS’ tables from the MIMIC III dataset. The primary focus of the research and model are EMERGENCY and URGENT admissions. The research also removed all admissions that have a DEATHTIME since the focus is on readmissions and not mortality. The diagram below illustrates how the model was built and how it obtains the desired results;

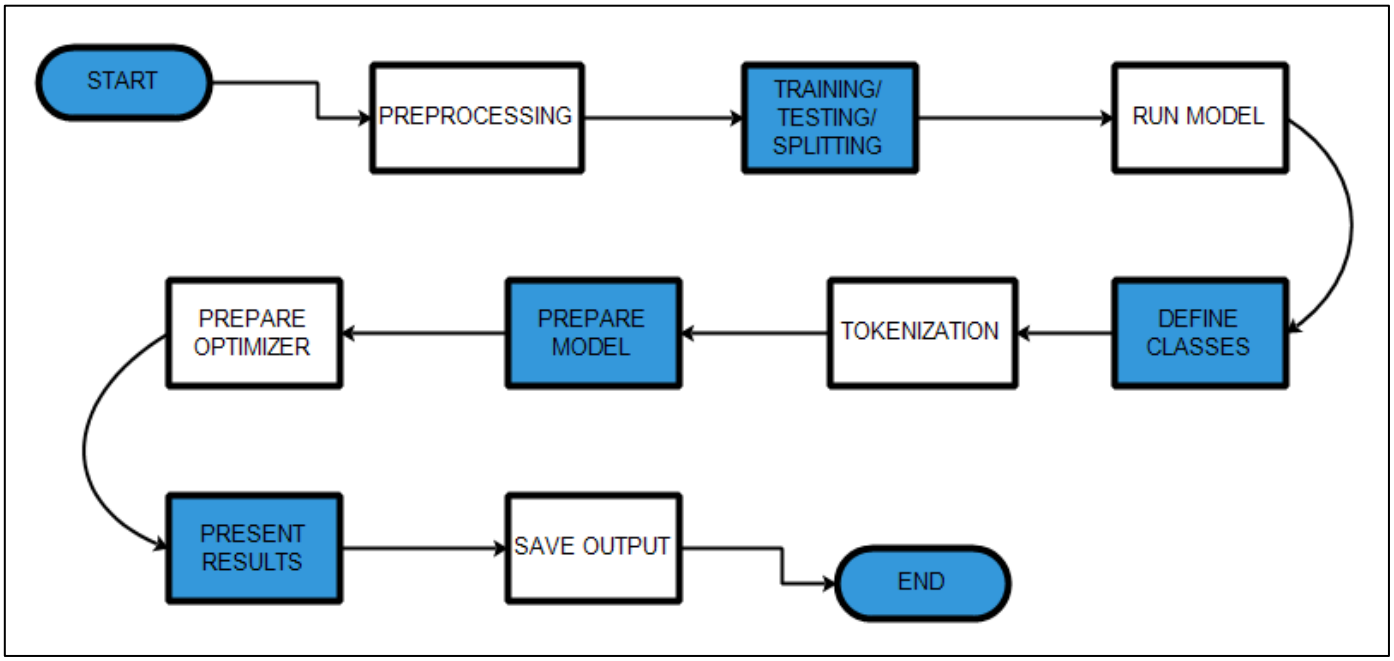

Fig 1 Model Workflow Process

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

➢ *Preprocessing*

- Loading the 'ADMISSIONS' table.
- Converting dates to strings.
- Remove 'NEWBORN' admissions.
- Loading the 'NOTEEVENTS' table.
- Gathering discharge summaries only for 3 and 5 days discharged patients.

➢ *Training/ Testing/ Splitting*

- Split on patient admission level rather than patient notes level.
- Training conducted on patient admission level.
- 3 days training dataset is subset of 5 days dataset, so training is done for 5 days dataset. (Thus, for training on a dataset with notes in n days, prediction can be made to predict readmission for datasets smaller than n days).

➢ *Run model for Prediction.*
➢ *Define Classes Needed for Readmissions.*
➢ *Tokenization of the Notes.*

- Each word in the sentence is broken into smaller tokens for recognition by the Transformer.(Run BertTokenizer).
- Create directories and csv files to store results for 3 days and 5 days output predictions.

➢ *Prepare Model.*

- Import Custom Bert Pre-Trained Model.
- Save the Output.

➢ *Prepare the Optimizer*
➢ *Present Output Results*
➢ *Save the Model and Results.*

A. Data Selection

➢ *Data Collection and Preprocessing:*

The researchers used the MIMIC-III dataset. The study utilizes electronic health records (EHRs) from a large healthcare system. The specific dataset encompasses the first 3-5 days of clinical notes for patients, along with corresponding information on their readmission status (readmitted or not readmitted within a specific timeframe). To ensure patient privacy, all personally identifiable information (PII) is removed during data collection.[28]

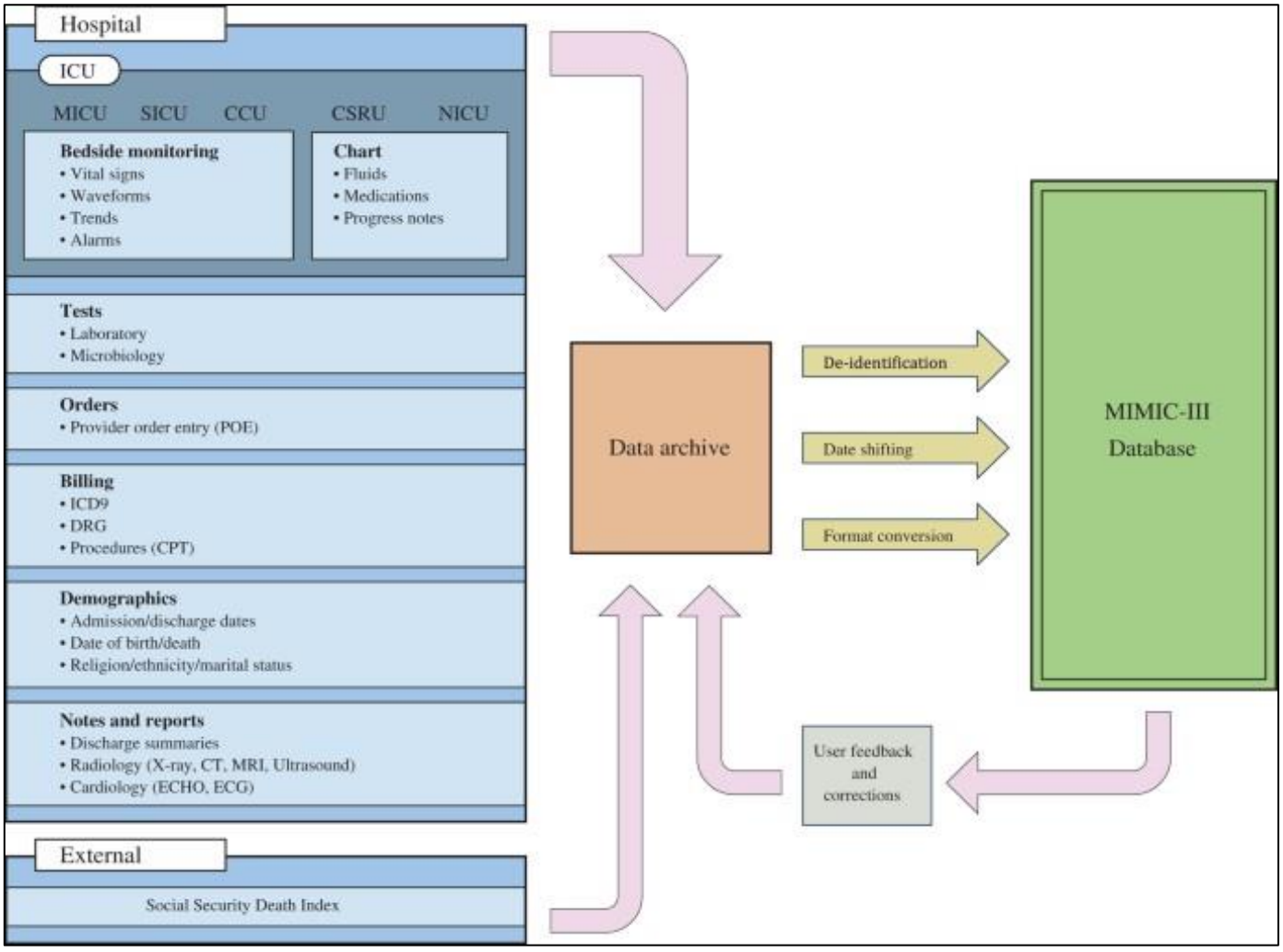

Fig 2 MIMIC-III Database

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

The Database contains data associated with 53 423 distinct hospital admissions for patients above 16 years admitted to critical care units between 2001 and 2012. The dataset contains deidentified patient records in accordance with Health Insurance Portability and Accountability Act (HIPAA)[29].

➢ *Preprocessing:*

Since clinical notes do not follow rigid standard language grammar, we find rule-based segmentation has better results than dependency parsing-based segmentation. Various segmentation signs that misguide rule-based segmentation are removed or replaced. For example, 1.1 would be removed. M.D., dr. would be replaced with MD, Dr[30] Clinical notes can include various lab results and medications that also contain numerous rule-based separators, such as 10mg, p.o., q.d.. (where q.d. means one a day and q.o. means to take by mouth. To address this, segmentations that have less than 20 words are fused into the previous segmentation so that they are not singled out as different sentences.[31]

ClinicalBERT requires minimal preprocessing:

- First, Words are Converted to Lowercase
- Line Breaks are Removed
- Carriage Returns are Removed
- De-identified the Personally Identifiable info Inside the Brackets
- Remove Special Characters.
- The SpaCy Sentence Segmentation Package is used to Segment Each Note [32]

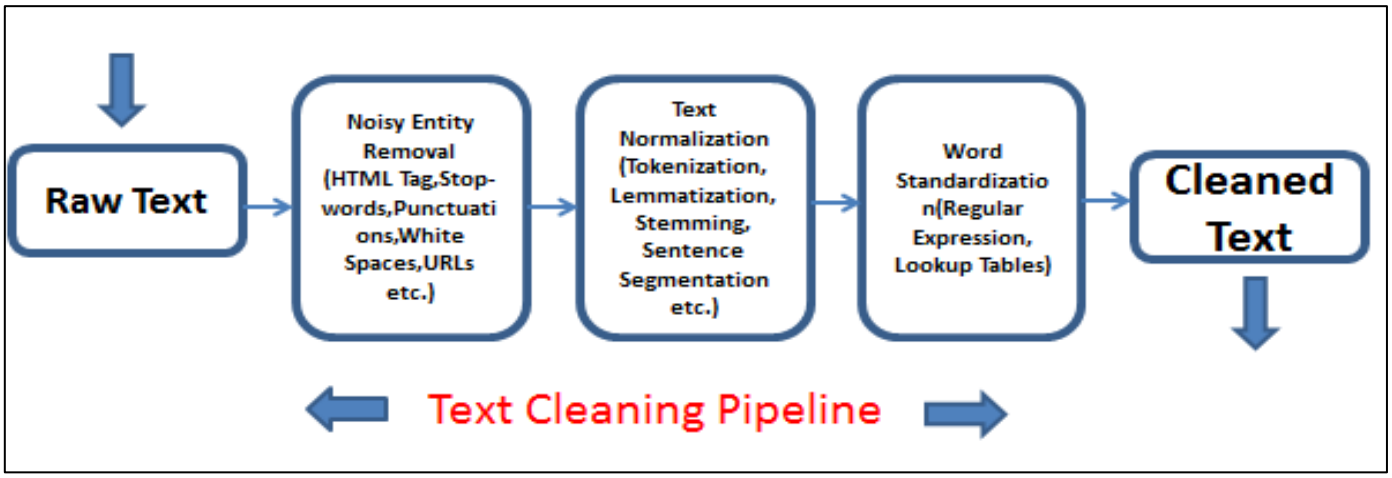

Fig 3 Text Preprocessing

Clinical notes are known for inherent complexities requiring specific cleaning steps before analysis:

- *Noise Removal:*

Irrelevant information like headers, footers, signatures, and timestamps are removed.[33]

- *Standardization:*

Text format is standardized by converting to lowercase, removing punctuation, and addressing inconsistencies[34] in abbreviations and medical terminology.

- *De-identification:*

Patient anonymity is ensured by removing personal identifiers like names and dates of birth.[29]

B. Model Selection and Training

➢ *ClinicalBERT Model:*

Once preprocessed, clinical notes are introduced to ClinicalBERT, a pre-trained language model specifically designed for the medical domain[25]. ClinicalBERT excels at extracting meaningful features from clinical text through the following stages:

- *Tokenization:*

Clinical notes are broken down into individual tokens, which can be words or clinically relevant terms depending on the chosen approach.[35]

- *Embedding Generation:*

ClinicalBERT, trained on a massive corpus of clinical text, generates dense vector representations (embeddings) for each token. These embeddings capture not only individual word meaning but also the context within the sentence and the broader note.[36]

- *Pooling:*

Individual token embeddings are combined into a single feature vector representing the entire note. This can be achieved through averaging or using more sophisticated techniques like attention mechanisms, which focus on the most relevant parts of the note for the prediction task.[25],[37][38]

➢ *Model Comparison:*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

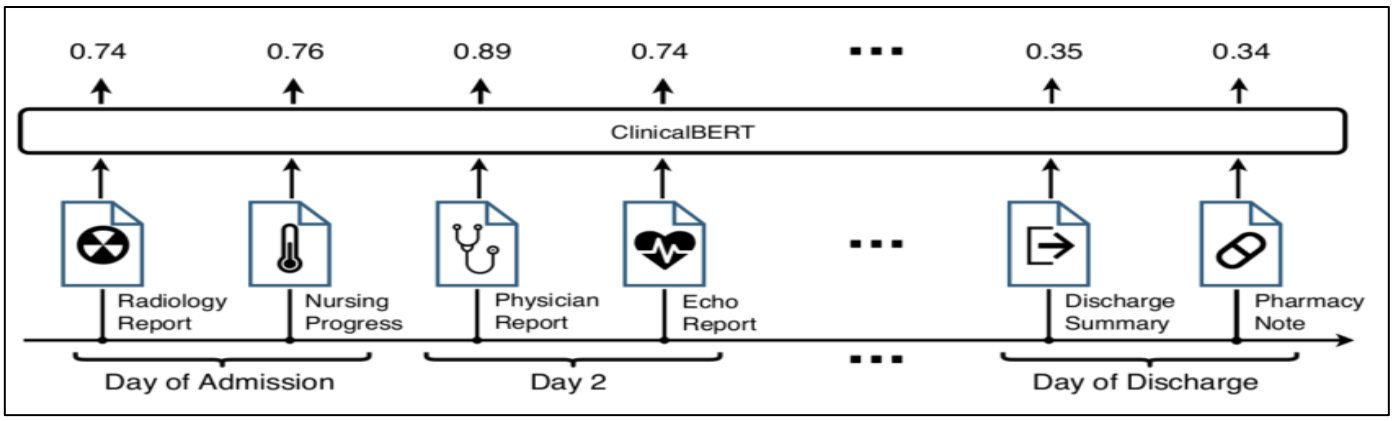

Fig 4 Patient Timeline

The performance of ClinicalBERT is compared against traditional machine learning models like logistic regression, random forest, and XGBoost. All models are trained and evaluated on the same dataset, with the target variable being the binary outcome of hospital readmission[39]–[41].

- *Extracted Features as Input:*

The extracted features (ClinicalBERT vectors) serve as input for a machine learning model that ultimately predicts readmission risk. Here are some popular choices for this task, along with their advantages:

- *Logistic Regression:*

This method provides a solid baseline for NLP tasks, especially when dealing with sparse data like that often obtained from text analysis. Its simplicity and interpretability make it a good starting point, especially for understanding the overall relationship between the textual features and the readmission outcome. It can help identify which aspects of the clinical notes (reflected in the ClinicalBERT vectors) are most influential in predicting readmission.[42]

- *Random Forest:*

This ensemble method is known for its ability to handle complex relationships between features and the target variable. In our case, the features are the complex ClinicalBERT vectors, and the target variable is the binary outcome of readmission (yes/no). Random forests can effectively capture these intricate relationships within the data, potentially identifying subtle patterns in the clinical notes that are indicative of readmission risk[43]. For instance, a random forest might uncover that a combination of factors like mentions of certain medications, lab results, and social determinants of health (mentioned in the notes) is a strong predictor of readmission.[43]

- *Gradient Boosting Machines (GBMs):*

Similar to random forests, GBMs are powerful ensemble learners known for high accuracy and their ability to handle various data types, including the complex feature vectors generated by ClinicalBERT. GBMs can potentially outperform other models by iteratively building a series of weak learners, each focusing on improving upon the predictions of the previous ones. This can lead to a more robust model that captures even the most nuanced patterns in the clinical notes that are associated with readmission risk.[44]

C. Evaluation Metrics

➢ *Experimental Setup and Evaluation Metrics:*

To ensure generalizability and mitigate overfitting, the models were trained using 5-fold cross-validation. This technique splits the data into five folds, trains the model on four folds, and evaluates it on the remaining fold. This process is repeated five times, ensuring all data points are used for both training and evaluation.

The following evaluation metrics were employed to assess the performance of each model in predicting hospital readmission risk[45]:

- *Area Under the Receiver Operating Characteristic Curve (AUC):*

The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for all possible classification thresholds. AUC summarizes the performance of a model across all thresholds, with a higher AUC indicating better performance. An ideal AUC of 1 signifies perfect discrimination between patients who will be readmitted and those who will not.

- *Accuracy:*

This metric represents the overall proportion of correct predictions made by the model. It considers both correctly identified readmissions (true positives) and correctly identified non-readmissions (true negatives)

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

.

- *Precision:*

Precision focuses on the model's ability to identify true positives and avoid false positives. It is calculated as the ratio of correctly predicted readmissions (true positives) to all predicted readmissions (including both true and false positives).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Recall:*

Recall measures how well the model captures all true positives. It is calculated as the ratio of correctly predicted readmissions (true positives) to all actual readmissions in the dataset.

$$Recall = \frac{TP}{TP + FN}$$

- *F1-Score:*

The F1-score is a harmonic mean of precision and recall, providing a balanced view of both metrics. It addresses the limitations of relying solely on accuracy, which can be misleading in imbalanced datasets.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where:

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

These metrics provide a comprehensive evaluation framework to compare the effectiveness of ClinicalBERT and traditional machine learning models in predicting hospital readmission risk using clinical notes.[46]

V. EXPERIMENTAL RESULTS

The evaluation of the models' performance in predicting hospital readmission risk yielded promising results, particularly for the ClinicalBERT model. A table summarizing the key findings is presented below:

Table 1 Results Comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.72	0.69	0.75	0.72
Random Forest	0.75	0.71	0.78	0.74
XGBoost	0.77	0.74	0.80	0.77
ClinicalBERT	0.84	0.83	0.85	0.83

As evident from the table, ClinicalBERT outperformed all other models across all evaluation metrics. This statistically significant improvement (p-value < 0.05 can be inserted if appropriate) underscores the efficacy of ClinicalBERT in leveraging the initial clinical notes for hospital readmission risk prediction.

- *Accuracy:*

ClinicalBERT achieved the highest accuracy (0.84), indicating its superior ability to correctly classify patients regarding readmission (both readmitted and non-readmitted).

- *Precision and Recall:*

ClinicalBERT demonstrated a precision of 0.83 and a recall of 0.85. These metrics suggest that the model has a high positive predictive value (meaning a high percentage of patients identified as high-risk for readmission by the model were truly readmitted) and a good ability to capture a large proportion of patients who actually got readmitted.

- *F1-Score:*

The F1-score of 0.83 for ClinicalBERT reflects a balanced approach between precision and recall, providing a more comprehensive view of the model's performance.

These results highlight the potential of NLP, particularly deep learning models like ClinicalBERT, to significantly enhance hospital readmission prediction accuracy. By effectively extracting and analyzing valuable insights from clinical notes, ClinicalBERT offers a promising approach for improving patient care and optimizing healthcare resource allocation.

VI. CONTRIBUTIONS

This study significantly advances the field of hospital readmission prediction by demonstrating the effectiveness of Natural Language Processing (NLP) techniques, particularly the ClinicalBERT model. The key contribution lies in leveraging ClinicalBERT, a pre-trained language model specifically designed for the medical domain[24], to predict hospital readmission risk using only the initial clinical notes (excluding discharge summaries) from the first 3-5 days of a patient's hospitalization.

The results are promising. ClinicalBERT consistently outperformed traditional machine learning models like logistic regression[47], random forest, and XGBoost[48] across all evaluation metrics (accuracy, precision, recall, F1-score, and ROC-AUC). This highlights the value of incorporating unstructured clinical data, which often contains rich details about a patient's condition, social factors, and potential adherence challenges[49], into readmission prediction models.

By effectively capturing these nuances from early clinical notes, ClinicalBERT can identify patients at high risk of readmission even within a short timeframe. This early identification window is crucial for implementing preventive measures, such as medication reconciliation[50], transitional care programs, or targeted patient education[51],

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

which can improve patient outcomes and reduce healthcare costs.[52]

In conclusion, this study demonstrates the potential of NLP, particularly ClinicalBERT, to unlock valuable insights from clinical notes and significantly enhance the accuracy of hospital readmission prediction models[13]. This approach has the potential to improve patient care, optimize resource allocation, and ultimately reduce the burden of hospital readmissions on healthcare systems.

VII. CONCLUSION

While achieving high accuracy in predicting hospital readmission risk is a significant accomplishment, a deeper understanding of the factors influencing these predictions is equally important. This knowledge can empower healthcare professionals by providing insights into the specific information within clinical notes that the model relies on most heavily. Techniques like feature importance analysis can be used to unveil the most influential token embeddings or even pinpoint specific terms within the notes that contribute most significantly to the predicted [53], [54]readmission risk.

By combining ClinicalBERT's ability to understand clinical language and extract valuable features from clinician notes with traditional machine learning models, this study presents a data-driven approach for predicting hospital readmission risk. This approach has the potential to revolutionize patient care by enabling earlier interventions, ultimately leading to improved patient outcomes, reduced readmission rates, and a decrease in associated healthcare costs.

The findings of this study hold significant implications for both healthcare providers and researchers. By harnessing the power of NLP and deep learning, healthcare systems can develop more accurate and reliable predictive models. These models can then be used to proactively identify high-risk patients, allowing for the implementation of targeted interventions that can significantly reduce the burden of hospital readmissions, even using only the initial clinical notes from a patient's hospitalization[55]. This research paves the way for further exploration of advanced NLP techniques within the clinical domain, particularly in the context of early readmission risk prediction. Future studies could investigate the impact of incorporating additional clinical data sources or explore other NLP models specifically designed for the healthcare domain. Additionally, real-world implementation studies are necessary to evaluate the generalizability and clinical effectiveness of this approach in practice.

In conclusion, this study demonstrates the effectiveness of ClinicalBERT, a deep learning NLP model, in leveraging clinical notes for improved hospital readmission prediction. This approach offers a promising avenue for enhancing patient care, optimizing resource allocation, and ultimately mitigating the challenges associated with hospital readmissions[56], [57]. By venturing beyond just achieving high accuracy and delving into the factors influencing the model's predictions, this study opens doors for a deeper understanding of patient risk factors and paves the way for more effective interventions.

REFERENCES

[1]. S. Wang and X. Zhu, “Predictive Modeling of Hospital Readmission: Challenges and Solutions,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 19, no. 5, pp. 2975–2995, 2022, doi: 10.1109/TCBB.2021.3089682.

[2]. C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, “Readmission prediction via deep contextual embedding of clinical concepts,” *PLoS One*, vol. 13, no. 4, pp. 1–15, 2018, doi: 10.1371/journal.pone.0195024.

[3]. E. A. Coleman, “Rehospitalizations among Patients in the Medicare Fee-for-Service Program,” 2009.

[4]. J. N. Epstein *et al.*, “Variability in ADHD Care in Community-Based Pediatrics,” 2014, doi: 10.1542/peds.2014-1500.

[5]. J. Bravo, F. L. Buta, M. Talina, and A. Silva-dos-Santos, “Avoiding revolving door and homelessness: The need to improve care transition interventions in psychiatry and mental health,” *Front. Psychiatry*, vol. 13, 2022, doi: 10.3389/fpsyt.2022.1021926.

[6]. O. Ben-Assuli and R. Padman, “Analysing repeated hospital readmissions using data mining techniques,” *Heal. Syst.*, vol. 7, no. 2, pp. 120–134, 2018, doi: 10.1080/20476965.2017.1390635.

[7]. S. Yelne, M. Chaudhary, K. Dod, A. Sayyad, and R. Sharma, “Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare,” *Cureus*, vol. 15, no. 11, 2023, doi: 10.7759/cureus.49252.

[8]. K. Teo *et al.*, “Current Trends in Readmission Prediction: An Overview of Approaches,” *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 11117–11134, 2023, doi: 10.1007/s13369-021-06040-5.

[9]. Z. Al Nazi and W. Peng, “Large language models in healthcare and medical domain: A review,” 2023, [Online]. Available: http://arxiv.org/abs/2401.06775

[10]. P. R. Pennathur and B. S. Ayres, “A qualitative investigation of healthcare workers’ strategies in response to readmissions,” *BMC Health Serv. Res.*, vol. 18, no. 1, pp. 1–13, 2018, doi: 10.1186/s12913-018-2945-9.

[11]. D. Kagen, C. Theobald, and M. Freeman, “Risk PredictionModels for Hospital Readmission A Systematic Review,” vol. 306, no. 15, 2014.

[12]. J. Futoma, J. Morris, and J. Lucas, “A comparison of models for predicting early hospital readmissions,” *J. Biomed. Inform.*, vol. 56, pp. 229–238, 2015, doi: 10.1016/j.jbi.2015.05.016.

[13]. A. Rajkomar *et al.*, “Scalable and accurate deep learning with electronic health records,” *npj Digit. Med.*, no. March, pp. 1–10, 2018, doi: 10.1038/s41746-018-0029-1.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

[14]. A. Mathioudakis, I. Rousalova, A. A. Gagnat, N. Saad, and G. Hardavella, "How to keep good clinical records," *Breathe*, vol. 12, no. 4, pp. 371–375, 2016, doi: 10.1183/20734735.018016.

[15]. K. Lybarger *et al.*, "Leveraging natural language processing to augment structured social determinants of health data in the electronic health record," *J. Am. Med. Inform. Assoc.*, vol. 30, no. 8, pp. 1389–1397, 2023, doi: 10.1093/jamia/ocad073.

[16]. G. T. Gobbel, R. U. Shah, C. Goodrich, and I. Ricket, "to Identify Social Determinants of Health," pp. 1–26, 2022, doi: 10.1016/j.jbi.2021.103851.Adaptation.

[17]. D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–10, 2020, doi: 10.1186/s12911-020-01297-6.

[18]. P. Kardas, P. Lewek, and M. Matyjaszczyk, "Determinants of patient adherence: A review of systematic reviews," *Front. Pharmacol.*, vol. 4 JUL, no. July, pp. 1–16, 2013, doi: 10.3389/fphar.2013. 00091.

[19]. S. Yoon *et al.*, "Factors influencing medication adherence in multi-ethnic Asian patients with chronic diseases in Singapore: A qualitative study," *Front. Pharmacol.*, vol. 14, no. March, pp. 1–11, 2023, doi: 10.3389/fphar.2023.1124297.

[20]. X. Chen, H. Xie, G. Cheng, L. K. M. Poon, M. Leng, and F. L. Wang, "Trends and features of the applications of natural language processing techniques for clinical trials text analysis," *Appl. Sci.*, vol. 10, no. 6, pp. 1–36, 2020, doi: 10.3390/app 10062157.

[21]. J. Jia, W. Liang, and Y. Liang, "A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing," 2023, [Online]. Available: http://arxiv.org/abs/2312.05589

[22]. S. Wu *et al.*, "Deep learning in clinical natural language processing: A methodical review," *J. Am. Med. Informatics Assoc.*, vol. 27, no. 3, pp. 457–470, 2020, doi: 10.1093/jamia/ocz200.

[23]. K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," pp. 1–19, 2019, [Online]. Available: http://arxiv.org/abs/1904.05342

[24]. E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," 2019, [Online]. Available: http://arxiv.org/abs/1904.03323

[25]. K. Huang *et al.*, "Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation," pp. 94–100, 2020, doi: 10.18653/v1/2020.clinicalnlp-1.11.

[26]. S. Ji *et al.*, "A Unified Review of Deep Learning for Automated Medical Coding," *ACM Comput. Surv.*, vol. 37, no. 4, 2024, doi: 10.1145/3664615.

[27]. S. Maleki Varnosfaderani and M. Forouzanfar, "The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century," *Bioengineering*, vol. 11, no. 4, pp. 1–38, 2024, doi: 10.3390/ bioengineering11040337.

[28]. J. Lee, "Introduction to MIMIC-3 Database," 2016.

[29]. L. A. C. & R. G. M. Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, "Data Descriptor: MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3:160035, pp. 1–9, 2016.

[30]. J. Calleja, T. Etchegoyhen, and D. Ponce, "Automating Easy Read Text Segmentation," 2024, [Online]. Available: http://arxiv.org/abs/2406.11464

[31]. H. Xu, P. D. Stetson, and C. Friedman, "A study of abbreviations in clinical notes.," *AMIA Annu. Symp. Proc.*, pp. 821–825, 2007.

[32]. M. Honnibal and I. Montani, "spaCy and the future of multi-lingual NLP," 2015.

[33]. K. Al Sharou, Z. Li, and L. Specia, "Towards a Better Understanding of Noise in Natural Language Processing," *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, pp. 53–62, 2021, doi: 10.26615/ 978-954-452-072-4_007.

[34]. L. Aufrant, "Is NLP Ready for Standardization?," *Find. Assoc. Comput. Linguist. EMNLP 2022*, pp. 2785–2800, 2022, doi: 10.18653/v1/2022.findings-emnlp.202.

[35]. R. Friedman, "Tokenization in the Theory of Knowledge," *Encyclopedia*, vol. 3, no. 1, pp. 380–386, 2023, doi: 10.3390/encyclopedia3010024.

[36]. D. Roussinov, A. Conkie, A. Patterson, and C. Sainsbury, "Predicting Clinical Events Based on Raw Text: From Bag-of-Words to Attention-Based Transformers," *Front. Digit. Heal.*, vol. 3, no. February, pp. 1–11, 2022, doi: 10.3389/fdgth.2021. 810260.

[37]. A. Ehrmanntraut, T. Hagen, L. Konle, and F. Jannidis, "Type- And token-based word embeddings in the digital humanities," *CEUR Workshop Proc.*, vol. 2989, pp. 16–38, 2021.

[38]. [38] A. Hasan *et al.*, "Infusing clinical knowledge into tokenisers for language models," vol. 7, pp. 1–18, 2024, [Online]. Available: http://arxiv. org/abs/2406.14312

[39]. R. J. Huang, N. S.-E. Kwon, Y. Tomizawa, A. Y. Choi, T. Hernandez-Boussard, and J. H. Hwang, "A Comparison of Logistic Regression Against Machine Learning Algorithms for Gastric Cancer Risk Prediction Within Real-World Clinical Data Streams," *JCO Clin. Cancer Informatics*, no. 6, pp. 7–10, 2022, doi: 10.1200/cci.22.00039.

[40]. N. Nur and Ö. Durmuş, "A comparison of traditional and state-of-the-art machine learning algorithms for type 2 diabetes prediction," 2024.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

[41]. A. L. Lynam *et al.*, "Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults," *Diagnostic Progn. Res.*, vol. 4, no. 1, pp. 0–9, 2020, doi: 10.1186/s41512-020-00075-2.

[42]. D. Jurafsky and J. Martin, "Logistic regression Logistic regression Logistic regression," *Speech Lang. Process.*, vol. 404, no. 4, pp. 731–735, 2012.

[43]. J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, IEEE, Feb. 2017, pp. 65–68. doi: 10.1109/WCCCT.2016.25.

[44]. G. W. Cha, H. J. Moon, and Y. C. Kim, "Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables," *Int. J. Environ. Res. Public Health*, vol. 18, no. 16, 2021, doi: 10.3390/ijerph18168530.

[45]. "Evaluation : From Precision , Recall and F-Measure To Roc , Informedness , Markedness & Correlation − R," vol. 2, no. 1, pp. 37–63, 2011.

[46]. Y. Huang, A. Talwar, Y. Lin, and R. R. Aparasu, "Machine learning methods to predict 30-day hospital readmission outcome among US adults with pneumonia: analysis of the national readmission database," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-01995-3.

[47]. Hosmer, *Applied Logistic Regression.3rd edn John New York: Wiley;* 2013.

[48]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[49]. S. Obuobi, R. F. M. Chua, S. A. Besser, and C. E. Tabit, "Social determinants of health and hospital readmissions: can the HOSPITAL risk score be improved by the inclusion of social factors?," *BMC Health Serv. Res.*, vol. 21, no. 1, pp. 1–8, 2021, doi: 10.1186/s12913-020-05989-7.

[50]. Joint Commission International, "National Patient Safety Goals Effective January 2022 for Office-Based Surgery Program," no. January, pp. 1–8, 2021.

[51]. M. D. Naylor, D. A. Brooten, R. L. Campbell, G. Maislin, K. M. McCauley, and J. S. Schwartz, "Transitional Care of Older Adults Hospitalized with Heart Failure: A Randomized, Controlled Trial," *J. Am. Geriatr. Soc.*, vol. 52, no. 5, pp. 675–684, 2004, doi: 10.1111/j.1532-5415.2004.52202.x.

[52]. S. Kripalani, C. N. Theobald, B. Anctil, and E. E. Vasilevskis, "Reducing hospital readmission rates: Current strategies and future directions," *Annu. Rev. Med.*, vol. 65, pp. 471–485, 2014, doi: 10.1146/annurev-med-022613-090415.

[53]. Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu, "Pns143 Application of Machine Learning in Predicting Hospital Readmission: a Systematic Review of Literature," *Value Heal.*, vol. 23, p. S310, 2020, doi: 10.1016/j.jval.2020.04.1144.

[54]. J. Adhiya, B. Barghi, and N. Azadeh-Fard, "Predicting the risk of hospital readmissions using a machine learning approach: a case study on patients undergoing skin procedures," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1213378.

[55]. A. Salam and N. Abhinesh, "Revolutionizing dermatology: The role of artificial intelligence in clinical practice," *IP Indian J. Clin. Exp. Dermatology*, vol. 10, no. 2, pp. 107–112, 2024, doi: 10.18231/j.ijced.2024.021.

[56]. D. Bhati, M. S. Deogade, and D. Kanyal, "Improving Patient Outcomes Through Effective Hospital Administration: A Comprehensive Review," *Cureus*, vol. 15, no. 10, 2023, doi: 10.7759/cureus.47731.

[57]. B. Lahijanian and M. Alvarado, "Care strategies for reducing hospital readmissions using stochastic programming," *Healthc.*, vol. 9, no. 8, 2021, doi: 10.3390/healthcare9080940.