Deep Learning Paper Review

Shauna Heron
Laurentian University

Table of contents

Introduction and Motivation

This review analyzes a study by R. Garriga, T. S. Buda, J. Guerreiro, J. O. Iglesias, I. E. Aguerri, and A. Matić [1], published in *Cell Medical*, titled *"Combining Clinical Notes with Structured Electronic Health Records Enhances the Prediction of Mental Health Crises."* The research investigates the utility of combining unstructured clinical notes with structured data from electronic health records (EHR) to improve the prediction of mental health crises [1].

The relevance of the research is underscored by a global rise in mental health-related hospitalizations coinciding with significant workforce challenges [2], [3]. Combined, these challenge stress the need for tools that might anticipate demand so that caseloads can be managed more efficiently. Even in Ontario, recent data highlights that hospitalizations for mental health conditions, particularly among youth, have surged since the COVID-19 pandemic [3], [2]. With those requiring hospitalization (e.g., emotional breakdowns, substance overdoses and suicide attempts) increasing by 90% [3]–many of which might have been prevented with early intervention, underscoring the idea that a tool that could predict the onset of mental health crises before they peak might conserve human resources as well as save lives [1].

Problem: modeling EHR data to predict crisis

As the authors point out, leveraging EHRs to bolster clinical decision-making is not new [1]. Clinicians and researchers have long utilized structured data like diagnosis codes, lab results, and medication records to inform predictive models [4]. However, computational limits as well as a lack of sophisticated modeling tools necessary to implement such models, meant that unstructured data like clinical notes and other free-form text, were most often left out of feature sets, despite the rich contextual information that clinical notes contain [1].

Solution: a hybrid model

With this gap in mind, the authors proposed a hybrid solution that built on a previous study which used structured EHRs to predict crisis events with a weekly crisis-risk score [5], this time adding unstructured data in the form of clinical notes to bolster those predictions (2023). Specifically, the author's hypothesized that models that combined both structured and unstructured data (such as clinical notes) would predict the probability of a weekly crisis event more accurately than a model including only structured data alone.

The structured data included static features like demographic information (e.g., postal code, race, gender), dynamic features like the most recent interactions with the hospital (e.g., substance use identified, new medication), and time-features that quantified the time that had elapsed since a specific event (e.g., time since the last crisis episode). Unstructured data was made up of free-form clinical notes written by clinicians. From these notes, semantic features were created using a BERT model, which is a pre-trained model good at extracting enhanced language features by considering the words that come before and after each word in a sentence [6].

Relying on those features four models were trained to predict a binary classification: relapse versus no relapse. The models included: i) a deep neural network trained on structured data (referred to as Struct DNN), ii) a model trained exclusively on unstructured data (Text DNN), iii) a hybrid model that integrated both structured and unstructured data (Hybrid DNN) and finally, iv) an ensemble of models that utilized predictions from the Hybrid DNN when unstructured data was available and predictions from the Struct DNN model otherwise (referred to as Ensemble DNN) was implemented. The experimental design and modeling process is outlined in Figure 1.

Model Architecture and Data Flow Diagram

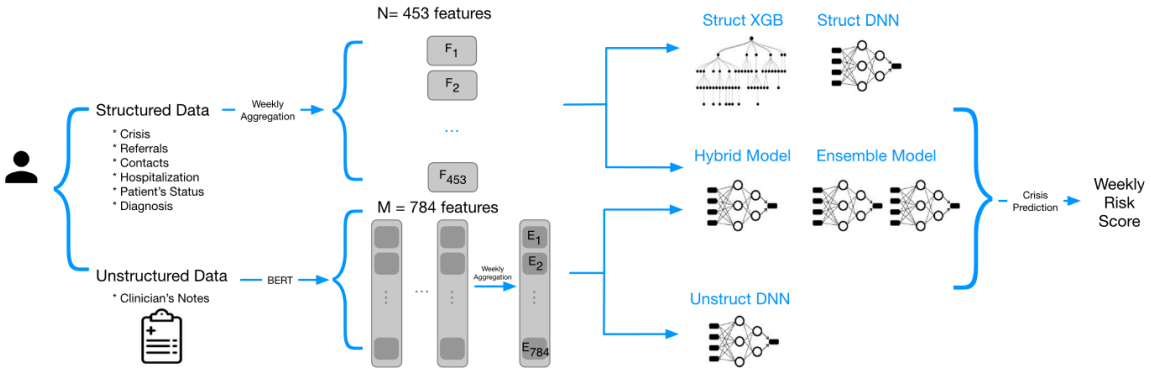

Figure 1: *Overview of the Hybrid DNN model and data input structure, showing how structured and unstructured data are processed. Adapted from [1, pg. 3.];*

To address class imbalance issues (since crisis relapses are rare with only 1.3% prevalence), models were tuned to maximize the area under the precision-recall curve (AUPRC)–which as we learned in class is a particularly useful metric when evaluating results produced by imbalanced classes [7]. Moreoever, as highlighted in our lessons, precision-recall metrics provide a better measure of

a model's predictive quality than accuracy in healthcare settings where false positives can be just as costly as false negatives [7], [1].

To evaluate model performance against non-machine learning methods used to evaluate crisis-risk, two baselines were built: a 5-factor logistic regression model informed by significant variables suggested in prior literature (LogReg5) and a heuristic model that ranked patients by the total number of crises they had experienced in the past year [1].

Results

The experiments revealed that the best structured-data-only model was an XGBoost classifier, a tree-based model implementing gradient boosting. The best performing model on unstructured data as well as combined structured and unstructured data was a feed-forward deep neural network (DNN) [5]. Both of these models performed better than both baseline models, but the best performing model overall was the ensemble DNN that utilized predictions from a structured DNN when unstructured data was unavailable and the hybrid DNN combining data types when it was, enabling the model to draw insights from semantic features when available to increase overall performance [5].

Importantly, the authors also examined the impact that various features in the models had on predictive performance using Shapley additive explanations (SHAP) [5], [8]. The authors use of SHAP values to assess the predictive power of different types of data as well as different features is significant because it increased model interpretability and transparency which is critical in healthcare settings were decisions need to be trusted by front line staff [9]. Though an in depth discussion of SHAP analysis, including its controversies [10], is beyond the scope of this review, it is worth noting that the authors implemented it; which allowed for some degree of both cohort-level and instance-specific explanations of predictions [8].

Performance Comparison Across Models

The comparative performance of the Structured-XGBoost, unstructured DNN and the hybrid DNN combining data types at predicting crisis in 10-week intervals across the 52 weeks are illustrated in the figures below.

According to the authors, the increase in AUPRC over time demonstrated in Figure 2 is only weakly related to an increase in EHR data over time, but more to the clinical fact that the chance of a crisis-relapse increases the longer a patient is in treatment [1]. The structured XGBoost model outperformed both models at 10-weeks, but at 20 weeks, when clinical note data becomes available, the XGBoost model is surpassed by the hybrid model. Higher AUPRC values indicate better precision in predicting true crisis-events relative to non-crisis events that are incorrectly predicted [11].

AUPRC for Predictions Across 52 Weeks

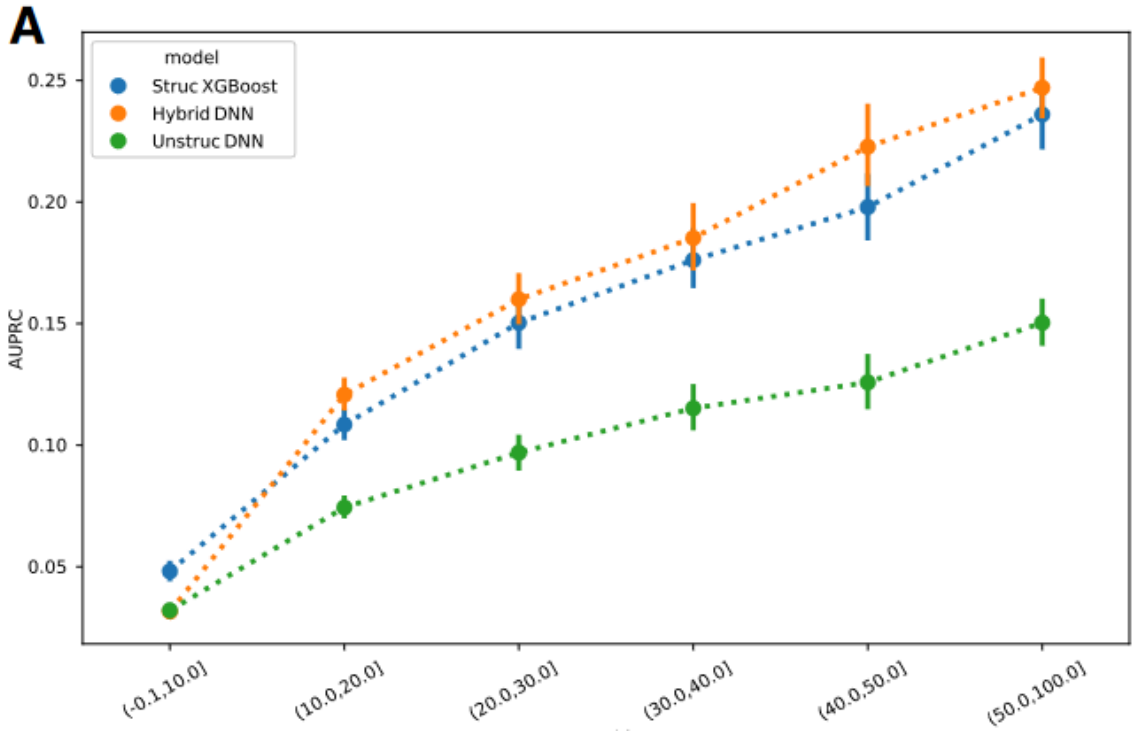

Figure 2: *Precision-recall curves for the Hybrid DNN, Struct DNN, and Text DNN, in10-week intervals over 52 weeks show an increase in performance across all models over time. Adapted from [1, pg. 5];*

The receiver operating characteristic (AUROC) comparisons in Figure 3 similarly reflect the model's ability to distinguish between true positive and negative predictions, however AUROC is considered a more forgiving metric when evaluating imbalanced datasets because it measures the model's overall capacity to separate classes, rather than focusing exclusively on predicting positive classes (crisis events) [12].

As indicated in Figure 3, the ability of both the structured XGBoost and unstructured DNN models to distinguish between classes *decreases* over time until 40 weeks when both the structured and hybrid model increase in performance. Interestingly, the performance of the unstructured-only model decreases across the entire 52 weeks, which the authors suggest highlights the complexity of modelling mental health crises: while the quantity of data in a given client file tends to increase over time, the overall number of patients requiring treatment *decreases* which impacts performance of the neural network (i.e., text input tends to require more examples or training instances to perform adequately in NNs) [1], [13].

Note too how the structured-only xgboost is best at predicting crisis with fewer than 10% of notes, but with 20% of weeks with available notes, the hybrid model's performance offers a small improvement on performance, indicating that the addition of unstructured data helped the model make more accurate predictions across a range of thresholds, improving its overall discriminatory power.

AUROC Scores for Predictions Across 52 weeks

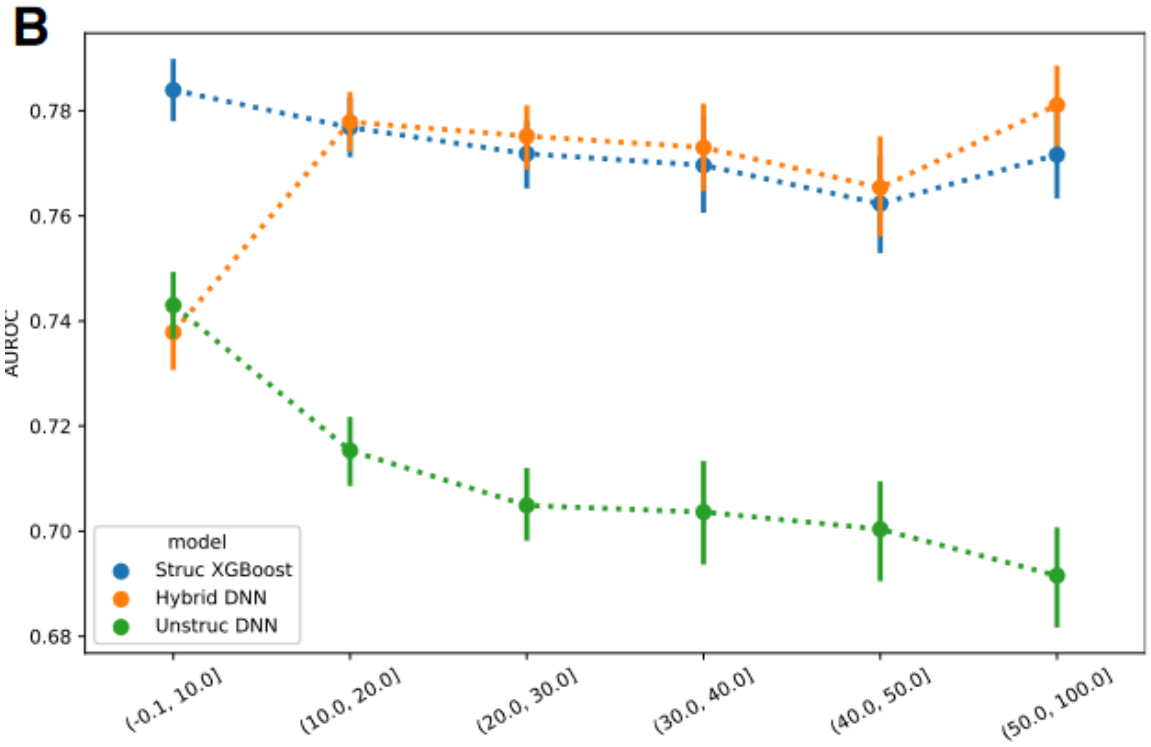

Figure 3: *Performance of the Hybrid model compared to the unstructured and structured only models in predicting crisis episodes at 10 week intervals. Points and lines indicate mean and ± standard deviation values computed in the 52 weeks of the test set. Adapted from [1, pg. 5];*

Importantly, the ensemble model produced a mean AUPRC of 0.133, which was above the baseline positive class rate of 1.3% or .013 for crisis relapse, indicating the models were better at predicting crisis onset than chance alone [1]. Moreoever, the AUROC scores averaged 0.87 for the best model which was significantly higher than the chance threshold for a binary classification at 0.5.

SHAP analysis found that while structured data features dominated the top individual predictors for mental health crisis prediction, unstructured data (such as clinical notes) carried substantial predictive value when combined [1]. Although no single unstructured feature was highly impactful on its own, together they contributed more significantly than structured data alone. SHAP also showed that the predictive power of both structured and unstructured data increased with the availability of clinical notes, highlighting the importance of including both data types for better model performance [1]. See Figure 4.

Contribution of data types for the hybrid DNN at 10 week intervals

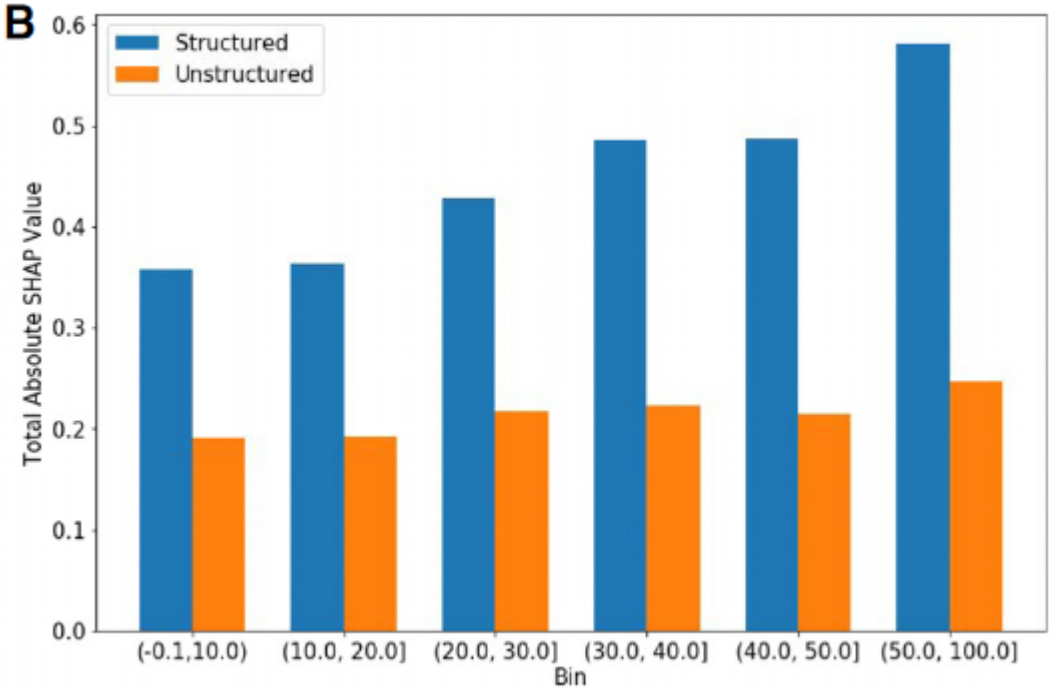

Figure 4: *The total absolute SHAP values for the Hybrid DNN for structured and unstructured feature categories extracted on the test set across the different datasets based on the percentage of notes available from the patients. Adapted from [1, pg. 5];*

Conclusion

In conclusion, the study effectively demonstrated that incorporating unstructured clinical notes alongside structured EHR data improved the ability to predict mental health crises. Importantly their findings underscore the potential for including qualitative clinical data in predictive models using a novel ensemble methodology that might be used to build flexible models that adapt based on the availability of data in the client record. Furthermore they demonstrated the utility of implementing feature analysis to increase interpretability and transparency of healthcare models. Future research could explore the generalizability of these findings to other domains within healthcare. In particular, it would be interesting to examine whether their methodology could be used to model EHR in community mental health settings where symptomology may not be as extreme and unstructured data types more varied. Another potential direction would be to try improving methods for natural language processing (NLP) to better extract meaningful insights from clinical notes, which might further enhance predictive performance.

Critical Analysis

All things considered, the paper offered a compelling case for the inclusion of unstructured data to bolster predictive modeling of healthcare data. The paper was particularly interesting to in the context of my own research where I aim to utilize structured EHR to predict case complexity in a community based mental health care agency. Importantly, they demonstrated and discussed sev-

eral less common techniques, including a flexible ensemble approach to modeling that adapts depending on data availability [1]. In addition, they shared an in-depth SHAP feature analysis which allowed both cohort-level and instance-specific analysis critical for enhancing transparency in healthcare machine learning models [4]. This was a fascinating component that I could have written a paper on all on its own.

While the paper was impressive overall, with many topics we could have explored further, the performance achieved by even the best model was modest at best. While the AUPRC and AUROC scores were notable in the context of predicting a rare event like a mental health crises, the results still reflect a gap between predictive performance and the reliability needed for clinical decision-making [4]. The scores tells us that while the model can identify crises better than random guessing, the model may still flag many non-crisis events that could potentially overburden clinicians [1]. A more thorough discussion of the model's threshold for clinical utility would have been good to see and is perhaps in the works. In their earlier study they conducted a post-hoc cohort study to examine clinical utility in everyday practice which would have been nice to see here too [5].

Overall the paper was very strong in terms of communicating the results in a way that was digestible and straightforward. The language was not full of jargon nor complicated equations and the supplementary materials were robust and detailed. Whether it was the Python code used for each stage of the model building process or more in depth explanation of methodology, including tables of predictions and outcome metrics–all were included in the supplementary materials. In terms of its reproducibility the paper was excellent and will prove extremely useful for anyone hoping to use their methodology as a springboard for future research–which I hope to do!

Bibliography

[1] R. Garriga, T. S. Buda, J. Guerreiro, J. O. Iglesias, I. E. Aguerri, and A. Matić, "Combining clinical notes with structured electronic health records enhances the prediction of mental health crises," *Cell Reports Medicine*, vol. 4, no. 11, Nov. 2023, doi: 10.1016/j.xcrm.2023.101260.

[2] N. Roumeliotis *et al.*, "Mental Health Hospitalizations in Canadian Children, Adolescents, and Young Adults Over the COVID-19 Pandemic," *JAMA Network Open*, vol. 7, no. 7, p. e2422833, Jul. 2024, doi: 10.1001/jamanetworkopen.2024.22833.

[3] CMHO, "Addressing Urgent Workforce Challenges in Child and Youth Mental Health," Mar. 2022.

[4] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 40–41, Aug. 2023, doi: 10.1186/s43067-023-00108-y.

[5] R. Garriga *et al.*, "Machine learning model to predict mental health crises from electronic health records," *Nature Medicine*, vol. 28, no. 6, pp. 1240–1248, Jun. 2022, doi: 10.1038/s41591-022-01811-5.

[6] "BERT 101 - State Of The Art NLP Model Explained." [Online]. Available: https://huggingface.co/blog/bert-101

[7] M. C. Lau, “Machine Learning / Deep Learning - Classifier Evaluation Slides.” 2024.

[8] “Welcome to the SHAP documentation — SHAP latest documentation.” [Online]. Available: https://shap.readthedocs.io/en/latest/

[9] C. C. Yang, “Explainable Artificial Intelligence for Predictive Modeling in Healthcare,” *Journal of Healthcare Informatics Research*, vol. 6, no. 2, p. 228–229, Feb. 2022, doi: 10.1007/s41666-022-00114-1.

[10] X. Huang and J. Marques-Silva, “On the failings of Shapley values for explainability,” *International Journal of Approximate Reasoning*, vol. 171, p. 109112–109113, Aug. 2024, doi: 10.1016/j.ijar.2023.109112.

[11] OpenAI, “CHatGPT (Version 4),” 2024, [Online]. Available: https://openai.com/

[12] “Measuring Performance: AUPRC and Average Precision.” [Online]. Available: https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/

[13] M. C. Lau, “Machine Learning / Deep Learning - Neural Network Slides.” 2024.