PROJECT PROPOSAL FOR A HYBRID ATTENTION-BASED DEEP LEARNING MODEL TO IMPROVE CREDIT RISK PREDICTION

Arnav Talwani
Student# 1008233662
arnav.talwani@mail.utoronto.ca

Dhanishika Antony Jotheeswaran
Student# 1010131073
d.antony@mail.utoronto.ca

Dharuniga Antony Jotheeswaran
Student# 1010102112
dharuniga.antony@mail.utoronto.ca

Vishwajith Subhashraj
Student# 1010009688
vishwa.subhashraj@mail.utoronto.ca

ABSTRACT

In this proposal, we outline a Hybrid Attention Network for credit risk prediction, which is designed for tabular financial data. Prior methods, including logistic regression, decision trees, SVMs, and ensembles, struggle with complex data and analysis, limiting real-world applications. Our model will integrate feature embeddings, multi-layer perceptrons (MLPs), and attention mechanisms to better capture complex nonlinear relationships compared to traditional models. We will process features such as income, loan amount, FICO score, and debt-to-income ratio through parallel MLP and attention branches before merging into a sigmoid classifier. Our planned experiments include varying the depth of the MLP and performing ablation studies on the attention modules with evaluation targeting AUC-ROC > 0.85 and improved precision-recall trade-offs. A logistic regression model will serve as the baseline. Ethical considerations such as bias, fairness, and privacy will be addressed through subgroup performance audits and explainability tools such as SHAP. Our goal is to assess whether deep learning can offer improved accuracy and transparency in credit risk decisions and can supplement existing tools to improve classification in all cases.
—-Total Pages: 8

1 INTRODUCTION

Through this project, our goal is to develop a deep learning model to assess credit based on individual financial and demographic data. Credit risk refers to the likelihood that a borrower will default on a loan, meaning they fail to repay the borrowed amount, and accurate risk assessment is essential for banks and financial institutions to make informed lending decisions. Traditionally, credit scoring models rely on statistical methods such as logistic regression, decision trees, or ensemble models like random forests, which use characteristics including income level, employment history, age, credit history, and existing debt to estimate a borrower's likelihood of repayment ((Trivedi, 2020)). Although these models are effective, they often rely on preset rules and statistics, making it more difficult to capture complex non-linear relationships between variables. These limitations become increasingly evident as the volume, diversity and complexity of the data increase, making it harder for the models to scale or adapt to new data sources and profiles. Deep learning models have done well in extracting hidden patterns from high-dimensional datasets, making them suitable for structured data-based credit risk analysis. As financial systems become more digitalized with various data sources, there is a growing opportunity to incorporate deep learning to assess complex patterns in borrowing behaviour.

However, deep learning also introduces some limitations and risks, such as the introduction of potential bias through demographic data and the reduced interpretability to explain the reasoning behind decisions to accept or reject a loan. Therefore, our aim is not to replace existing risk assessment tools, but to investigate whether deep learning models can be used to supplement traditional methods to improve classification, especially in edge cases, such as underrepresented populations, where it is difficult for standard models to make predictions.

The motivation for this approach is to explore how deep learning can be responsibly used for complex financial decision-making. With an efficient and well-designed model, there can be reduced default rates, better loan allocation, and security for financial institutions. It will also reduce the time, effort, and subjective judgment currently required by analysts and credit assessors. A well-trained model can automate a significant portion of the initial assessment process, allowing employees to focus on more risky cases that require expert judgment, improving their operational efficiency. This proposal outlines the project's background, data processing, model architecture, and training and validation strategy to form a comprehensive approach to investigating credit risk assessment with deep learning.

2 Background & Related Work

2.1 Binary Classifiers for Loan Default Prediction

This study explored common machine learning methods like logistic regression and decision trees to predict whether people would repay loans ((Addo et al., 2018)). While models like decision tress are simple and more efficient, they often struggle when data is complex and includes many interacting features. In this study, they selected the 10 most important features from the model which introduced more abstraction. Also, most of these models only predict two categories: ”good” or ”bad” credit.

2.2 Deep Learning for Structural and Reduced-Form Credit Risk Models

Manzo & Qiao (2021) combined deep learning with the unscented Kalman filter, a technique for non-linear state estimation, to calibrate credit risk models. They were able to achieve an in-sample R-squared of 98.5 % for the reduced-form model and 95 % for the structural model. However, their model treated all input features equally and it was difficult to determine which ones mattered most. This lack of interpretability is a major limitation for real-world adoption, since financial institutions require transparent reasoning behind approvals and denials to meet regulatory standards. This makes organizations hesitant to adopt machine learning in credit risk settings despite its success.

2.3 Comparing ML Models and Challenge of Black-Box Interpretability

A study by S&P Global showed that machine learning models like Support Vector Machine (SVM), logistic regression, and decision trees have stronger performance in credit risk prediction compared traditional statistical methods ((Vidovic & Yue, 2020)). These models can capture complex, non-linear relationships and interactions between variables, making them valuable for identifying subtle risk patterns in applicant data. However, models like SVM use nonlinear kernel functions which make it a black box. This makes it difficult to attribute a prediction to a feature, reducing interpretability similar to Manzo & Qiao (2020). This study focused on private companies, and they only used features that a wide range of data in their sample. However, this is different from real-life applications which we want to address in our model.

2.4 Deep Learning Ensembles with Improved SMOTE for Imbalanced Data

Shen et al. (2021) combined LSTM networks with AdaBoost to develop a deep learning ensemble model that addressed imbalanced credit datasets. They improved classification performance by developing an improved synthetic minority oversampling technique (SMOTE) method. This was done using a Mahalanobis distance-based oversampling (MDO) technique which generates synthetic samples without class decomposition.

2.5 Stacked Classifiers with Filter-Based Feature Selection

This study introduced a stacked classifier approach that combines Random Forest, Gradient Boosting, and XGBoost for credit risk prediction across multiple datasets ((Emmanuel et al., 2024)). By applying filter-based feature selection using Information Gain, the model improved generalization and achieved high Area Under the Curve (AUC) scores (0.934, 0.944 and 0.870). However, the sequential structure of the ensemble adds complexity and reduces interpretability, making it difficult to trace how specific features influence predictions. Feature selection, although useful for reducing complexing and overfitting, can lose some subtle interactions between features.

2.6 Our Contribution: Binary Risk Prediction with Group-Wise Attention

While prior research has demonstrated the success traditional and advanced machine learning models for credit scoring, our project addresses some of the limitations identified across these studies. Addo et al. (2018) and Vidovic & Yue (2020) showed that logistic regression, decision trees, and SVMs have high accuracy but struggle with complex interactions and lack interpretability. Manzo & Qiao (2021) introduced deep learning with the unscented Kalman filter, which had a strong performance, but their model treated all features equally, making it difficult to explain decisions. Similarly, Shen et al. (2021) and Emmanuel et al. (2024) used ensemble models and feature selection to improve accuracy, yet their techniques introduced complexity, and reduced features and transparency. Our project builds on these insights by using a binary classification system with group-wise attention to a publicly available dataset (Kaggle). Group-wise attention puts related features are grouped, like financial info or personal demographics, together which allows the model to not only learn which features matter, but also how groups of features interact. Although our approach may not achieve the same accuracy as more complex networks, it can improve interpretability, which is crucial for application in financial institutions. We can also preserve more nuanced interactions between features, while improving the efficiency of training through our use of grouping.

3 Data Processing

We will be using the credit risk analysis dataset from Kaggle, which contains 855,969 loan records issued between 2007 and 2015 ((Mehta, 2020)). The dataset includes 73 features, including loan amount, interest, term, annual, income, and grade. The target variable is binary, indicating whether a loan defaulted (1) or not (0). Default is defined as a borrower failing to make timely payments, missing payments, or stops making payments. The dataset is highly imbalanced, with only about 6% of loans being defaulted.

Initial data processing will involve removing entries with missing values in key features, and removing columns that have irrelevant information (such as ID numbers). Numerical features like loan amount, interest rate, annual income, and installments will be standardized using z-score normalization. Categorical features, such as grade, purpose, and term, will be transformed using one-hot encoding. We will also organize the input features into meaningful groups, such as loan information, borrower information, and credit history. These groupings will be used in our group-wise attention model to create more interpretable and efficient learning, allowing us to determine which feature categories contributed most to a prediction.

Due to the severe class imbalance, we will randomly under sample the majority class (non-default loans) during training. This will ensure the model will be exposed to a balanced set of default and non-default cases. The dataset will also be split into training (70%), validation (15%), and test (15%) sets, preserving class distribution in each split.

4 Architecture

The proposed model predicts credit risk using a Hybrid Attention Network tailored for tabular financial data. This architecture integrates feature embedding, multi-layer perceptrons (MLPs), and attention mechanisms to learn the weights of various features such as loan amount, term and interest rate.

The model processes raw inputs through parallel paths: an attention branch and an MLP branch, which are used for identifying critical relationships and handling numerical features like annual_income with ReLU and dropout, respectively. These paths merge into a sigmoid output, yielding default probabilities. The team has thought of a couple of experiments, including testing the impact of MLP depth (such as 3 vs. 5 layers) and running an ablation study to isolate attention's contribution (targeting a +5% AUC gain over simpler designs). To ensure robustness, the architecture is evaluated against a logistic regression baseline, the industry standard for credit scoring. The hybrid design's performance is measured via AUC-ROC (targeting > 0.85) and precision-recall trade-offs. The goal is to minimize false approvals (precision) and capture true defaults (recall). This approach not only addresses the limitations of linear models but also provides transparency through attention weights. This will enable lenders to understand how specific features drive risk predictions

5 ILLUSTRATION

Figure 1 provides a high level depiction of the proposed model for credit risk prediction:

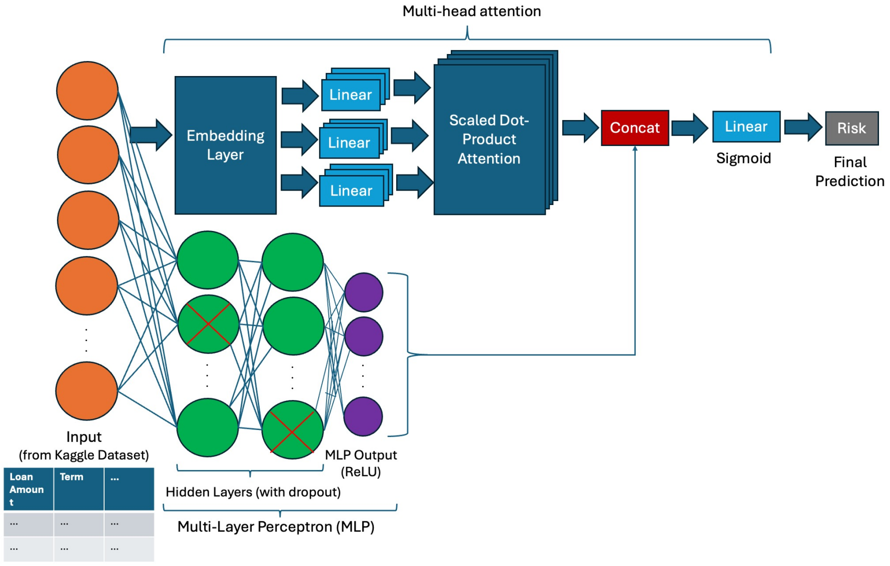

Figure 1: A high level illustration of the proposed architecture and organization of the overall model, detailing the input of the dataset, a high-level breakdown of the Hybrid Attention Network consisting of both a Multi-Head Attention stream and a Multi-Layer Perceptron stream, and the appropriate output of credit risk probability. Note that the "Linear" layers in the multi-head attention stream are fully-connected layers as depicted in the MLP stream, but have been abstracted to single layers due to depict the sheer amount of them used for multi-head attention.

6 BASELINE MODEL

The team decided to use Logistic Regression (LR) as the baseline model since it aligns with industry standards for credit risk assessment, and its probabilistic output makes it suitable for binary classification of loan defaults. The model will use predictors such as debt-to-income ratio, FICO score, and loan amount, after normalization and cleaning. The target variable is binary: "Default" (1) versus "Not Default" (0). Training employs L2 regularization (with $C = 1.0$) and class weighting to address data imbalance (approximately 6% defaults). Performance is evaluated using AUC-ROC (target ≥ 0.75), with precision and recall as secondary metrics. This baseline model provides a conservative benchmark, and its coefficients also offer transparency into feature importance, which the team will analyze in depth. Hence, the failure of the DL model to surpass LR would indicate architectural or data issues.

"

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(class_weight='balanced', max_iter=1000)
lr.fit(X_train, y_train)
```

Listing 1: Simplified training snippet (Scikit-learn)

7 Ethical Considerations

As with any Deep Learning project, there are a few ethical considerations to keep in mind. Firstly, there is a risk of historical biases such as higher interest rates for low-income ZIP codes or ethnic enclaves victims of redlining to propagate. We can mitigate this by auditing the model performance by subgroups, such as homeowners vs.renters. Privacy is also a concern since even anonymized data like this could be used to re-identify applicants. We could apply K-anonymity but with so many features (e.g., state, income, loan_amount), it would require excessive generalization, thereby reducing data utility. We could apply light generalization, such as by sorting income into $10K ranges, aggregating states into regions, and using differential privacy. Another big consideration lies in the impact of our results. False negatives (missed defaults) result in financial losses for lenders, while false positives (unfair denials) may unjustly exclude qualified applicants. The team aims to address this by using SHAP values to provide clear explanations for denials, such as identifying high debt-to-income ratios as the primary reason for a loan denial.

8 Project Plan

The team will communicate through a Discord groupchat and will meet online (time permitting) to distribute work and finalize project documents through this platform. If an in-person meeting is necessary, one will be coordinated to accommodate everyone's availability to maximize productivity. There will be an internal deadline of two days before each deliverable is due for a first draft, and the group will finalize each deliverable collectively before submitting it on the night of the due date. Work for a deliverable will be split up at least a week before the internal due date, leaving one week for everyone to work on it, although this will vary based on the technical complexity and estimated time required to complete each deliverable. The points assigned to each deliverable task in the rubric of every assignment will be used as a guideline for even distribution of work. Figure 2 illustrates the complete timeline of project deliverables throughout the semester, and Table 1 lists the preliminary distribution of work throughout the project deliverables.

9 Risk Register

See Table 2.

10 Link to Github

Link:GitHub Repository

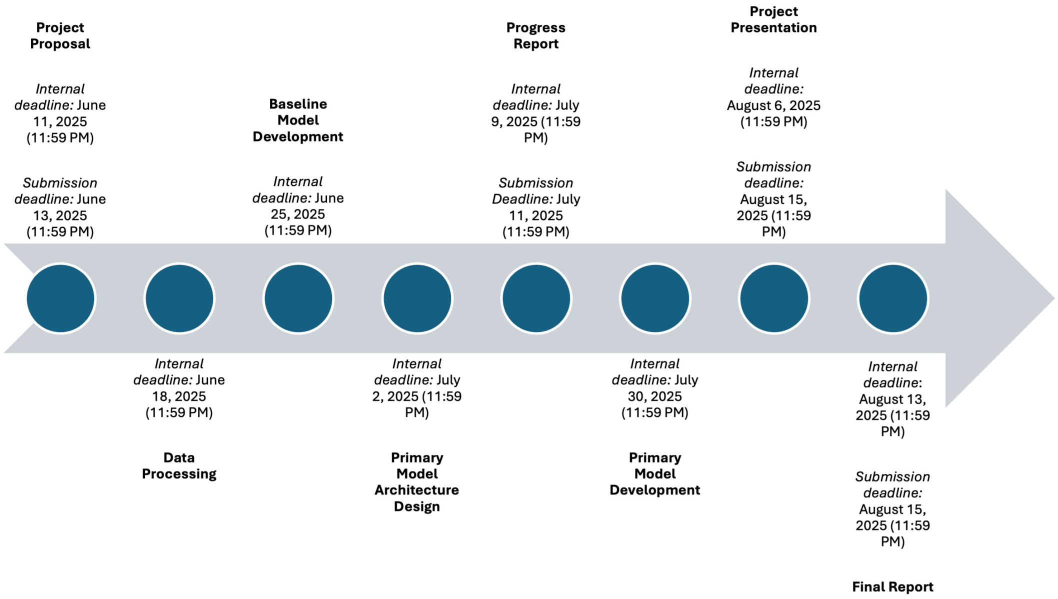
Project Proposal
Internal deadline: June 11, 2025 (11:59 PM)
Submission deadline: June 13, 2025 (11:59 PM)
Internal deadline: June 18, 2025 (11:59 PM)
Data Processing
Baseline Model Development
Internal deadline: June 25, 2025 (11:59 PM)
Internal deadline: July 2, 2025 (11:59 PM)
Primary Model Architecture Design
Progress Report
Internal deadline: July 9, 2025 (11:59 PM)
Submission Deadline: July 11, 2025 (11:59 PM)
Internal deadline: July 30, 2025 (11:59 PM)
Primary Model Development
Project Presentation
Internal deadline: August 6, 2025 (11:59 PM)
Submission deadline: August 15, 2025 (11:59 PM)
Internal deadline: August 13, 2025 (11:59 PM)
Submission deadline: August 15, 2025 (11:59 PM)
Final Report

Figure 2: A high level timeline of each of the project deliverables, including their internal and class (”official”) due dates

Table 1: Distribution of workload among team members for each project deliverable listed in Fig. 2

	Dhanishika	**Dharuniga**	**Vishwa**	**Arnav**
Project Proposal	Introduction Risk Register	Background Data Processing	Architecture Baseline Model Ethical Considerations	Visual Overview Project Plan
Data Processing	Data Collection and cleaning			
Baseline Model Development	Development of model architecture	Development and execution of training	Development and execution of testing	Development and execution of validation
Primary model architecture design	More detailed development of model architecture and specifics (TBD as development progresses)			
Progress Report	Project description Individual contributions	Data Processing write-up	Baseline Model write-up	Primary model write-up
Primary model development	Model architecture development	Development and execution of training	Development and execution of testing	Development and execution of validation
Project Presentation	Problem, data processing	Model	Results	Discussion
Final Report	Introduction Data processing Architecture	Illustration Background Baseline model	Quantitative and qualitative results	Discussion Ethical considerations

Table 2: Risk assessment and mitigation plan

RISK	DESCRIPTION	LIKELIHOOD	IMPACT	SEVERITY	RESPONSE
Teammate drops the course	A team member may withdraw due to personal or academic issues, increasing workload for the remaining members.	Low	High	Medium	Conduct regular check-ins via group chat, encourage open communication to identify issues early, and redistribute tasks accordingly.
Model training takes longer than expected	Model training may take longer due to computational constraints or complex data.	High	Low	Low	Monitor training progress closely, optimize model architecture and hyperparameters early, and adjust project timeline if necessary.
Unexpected bugs in model code	Unforeseen coding bugs may cause delays during model development and evaluation.	Medium	Medium	Medium	Allocate buffer time in project schedule, conduct regular code reviews, and implement systematic debugging and version control practices.
Labelling inconsistencies in the dataset	The dataset may contain mislabeled or inconsistent data, negatively impacting model performance.	High	Low	Low	Perform dataset review and cleaning, validate labels with cross-checks, and supplement dataset with additional reliable sources if required.
Difficulty in tuning hyperparameters	Hyperparameter tuning may take longer than expected or fail to yield optimal performance.	High	Medium	Medium	Use automated hyperparameter search methods (e.g., grid search, random search) and monitor validation performance.

References

Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. Credit risk analysis using machine and deep learning models. *Risks*, 6(2), 2018. ISSN 2227-9091. doi: 10.3390/risks6020038. URL https://www.mdpi.com/2227-9091/6/2/38.

Ileberi Emmanuel, Yanxia Sun, and Zenghui Wang. A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data*, 11(1):23, 2024. ISSN 2196-1115. doi: 10.1186/s40537-024-00882-0. URL https://doi.org/10.1186/s40537-024-00882-0.

Gerardo Manzo and Xiao Qiao. Deep learning credit risk modeling. *The Journal of Fixed Income*, 31:jfi.2021.1.121, 08 2021. doi: 10.3905/jfi.2021.1.121. URL https://openreview.net/pdf?id=F88KOHKRY3.

Ramesh Mehta. Credit risk analysis dataset. https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis, 2020. Accessed: 2025-06-12.

Feng Shen, Xingchao Zhao, Gang Kou, and Fawaz E. Alsaadi. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, 98:106852, 2021. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106852. URL https://www.sciencedirect.com/science/article/pii/S1568494620307900.

Shrawan K. Trivedi. A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, 63:101413, 2020. URL https://www.sciencedirect.com/science/article/pii/S0160791X17302324. ID: 271744.

Luka Vidovic and Lei Yue. Machine learning and credit risk modelling. Technical report, S&P Global, 2020. URL https://www.spglobal.com/marketintelligence/en/documents/machine_learning_and_credit_risk_modelling_november_2020.pdf.