CT6039 Dissertation

Student Name: Marcus Abraham

Student Number: s4203822

Course: Computer Science

Submission Date: XX XXX 2024

University of Gloucestershire

**Title goes here**

*One of these???*

*“Leveraging Ensemble Learning in Single Label Prediction of Adverse Reaction to Chemical Compounds”*

*“Leveraging Hybrid Architecture to Enhance Accuracy in Predicting Adverse Reactions to Preclinical Drugs”*

*"Ensemble Learning for Accurate Prediction of Adverse Reactions to Preclinical Drugs: A Comparative Study"*

*"Harnessing the Power of Ensemble Learning in Predicting Adverse Drug Reactions during Preclinical Trials"*

*"Improving Predictive Performance through Ensemble Learning: A Study on Adverse Reactions to Preclinical Drugs"*

*"A Comparative Analysis of Ensemble Learning Techniques for Predicting Adverse Reactions to Preclinical Drugs"*

*"Leveraging Ensemble Learning to Enhance Accuracy in Predicting Adverse Reactions to Preclinical Drugs"*

**Contents**

[**1. Abstract** 2](#_Toc138356804)

[**2. Introduction** 2](#_Toc138356805)

[**3. Literature Review** 3](#_Toc138356806)

[**4. Methodology** 3](#_Toc138356807)

[**4.1 Dataset Collection** 3](#_Toc138356808)

[**4.2 Feature Selection** 3](#_Toc138356809)

[**4.3 Techniques** 3](#_Toc138356810)

[**4.4 Evaluation Metrics** 3](#_Toc138356811)

[**5. Data Pre-processing & Analysis** 3](#_Toc138356812)

[**5.1 Data Pre-processing** 3](#_Toc138356813)

[**5.2 Data Analysis** 3](#_Toc138356814)

[**6. Classic Model Development** 3](#_Toc138356815)

[**7. Hybrid Model Development** 4](#_Toc138356816)

[**8. Evaluation** 4](#_Toc138356817)

[**9. Conclusion** 4](#_Toc138356818)

[**10. References** 4](#_Toc138356819)

[**11. Appendix** 4](#_Toc138356820)

# **1. Abstract**

An overview of what the paper contains

# **2. Introduction**

-Start with an overview of the issue presented by ADR’s, and how classical solutions struggle with catching side effects before expensive clinical trials. Highlight both the risks to human life, and costs for companies performing the trials that could have found side effects to developing drugs much sooner.

-State the objectives that this dissertation aims to achieve.

-List the research questions that need addressing.

-I will need to choose between single or multi label classification of ADRs. Either I’m making a model which can predict for one specific ADR, or I’m trying to make a model which can predict for all ADR’s a compound may have.

Problem statement:

How to increase accuracy of machine learning algorithms for drug side effect prediction, to reduce patient risk in clinical trials?

Research questions:

* What properties of a compound most influences its side effects?
* What are the current algorithms used, and how can they be improved?
* How effectively can a model be trained on public datasets?

Objectives:

* To investigate current machine learning techniques for drug side effect predict.
* To analyse and identify the most important features of compounds which influence their side effects.
* To develop and evaluate a new hybrid machine learning algorithm to enhance the accuracy of drug side effect predictions.

# **3. Literature Review**

-Gather a mix of around 10 papers/case studies for the review.

-Identify key gaps in the existing research.

# **4. Methodology**

## **4.1 Dataset Collection**

-Describe my datasets and other potential sources of ADR data.

## **4.2 Feature Selection**

-Explain which features I’m selecting from my combined datasets, and why. Cite papers that specifically tie these features to ADR’s.

## **4.3 Techniques**

-Explain each of the techniques and models I’m using in the project.

## **4.4 Evaluation Metrics**

-Explain how I am going to evaluate the success and accuracy of the model.

# **5. Data Pre-processing & Analysis**

## **5.1 Data Pre-processing**

-Go through how I cleaned, merged and prepared the datasets for analysis. Provide relevant screenshots and link papers which use the same analysis tools I do.

## **5.2 Data Analysis**

-Give statistics and visual analysis of the data.

-Describe any further feature engineering done.

# **6. Classic Model Development**

-Create 3 or 4 of the models from the literature review, the ones that got the best results. Fit them to my data, and see how they perform.

-For training, use five-fold cross-validation method. This means that the data is split into five parts, and the model is trained and validated on different combinations of these parts. This helps to find the best hyperparameters for the model.

-Evaluate their performance via accuracy, precision, recall, F1 score, k-fold cross-validation and other metrics laid out in the methodology.

# **7. Hybrid Model Development**

-Explain the design and architecture of my model.

-Detail the models training process, hyperparameter tuning and cross validation techniques used.

-Evaluate the models performance via its accuracy, precision, recall, F1 score, k-fold cross-validation and other metrics laid out in the methodology.

# **8. Evaluation**

-Analyse the results I got from my model, and compare them to the existing classical models.

-Discuss improvements or limitations that might have held my model back.

-Predict severity of ADRs

-Include patient data for enhanced accuracy

-Multi-label prediction rather than single-label

-Answer the research questions I wrote in the introduction.

# **9. Conclusion**

-Summarise my findings in the project, and how it has contributed to ADR prediction.

-Identify future improvements to my work, and other potential areas.

# **10. References**

# **11. Appendix**