



Story Point Estimation with Explainable AI

D M S S Dassanayake
(Reg. No.: MS23465190)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY SPECIALIZED IN
ENTERPRISE APPLICATION DEVELOPMENT

SEPTEMBER 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science

MSc Supervisor: Dr Mahima Weerasinghe

Approved for MSc. Research Project:

MSc in IT Program Coordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not that of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: 

D M Shehan Sandeepa Dassanayake

Date: 2025/11/03

ABSTRACT

Estimating Story points with Explainable AI

Dasswaththe Mudiyanseelage Shehan Sandeepa Dassanayake

MSc. In Information Technology Specialized in Enterprise Application Development

Supervisor: Dr. Mahima Weerasinghe

September 2025

Accurate story point estimation remains as a challenge in Agile software development because story point estimation depends on human intuition, experience, and subjectivity. Traditional story point estimation methods like planning poker and expert judgment most of the time lead to inconsistencies and biases, and it could impact resource allocation of the project and predictability of the project. This study addresses those limitations by seamless integration of transformer based natural language processing (NLP) models (BERT variants) and Explainable AI techniques (XAI) to interpret the estimation of story points.

Four transformer-based models, BERT-base-uncased, DistilBERT, RoBERTa-base, and DistilRoBERTa-base were trained using TAWOS dataset with baseline and customized preprocessing pipelines. Advanced preprocessing techniques such as adaptive Fibonacci mapping, semantic T5 based data augmentation, and context injection, improved the model accuracy. The model trained with DistilBERT achieved the highest performance with 0.520 with 10 early stopping patience and 0.507 accuracy with early stopping patience 4 and the lowest mean absolute error 0.72 (MAE = 0.72).

To improve the transparency and interpretability of trained models, XAI methods like SHAP and LIME explanations were applied. A survey with 8 agile practitioners showed strong alignment between XAI explanations and human explanations, with SHAP explanation achieving 72% and LIME explanation achieving 65% overlapping with agile practitioners identified keywords.

The findings show that transformer models with XAI achieved accuracy comparable to human estimations (80% agreement) with interpretable predictions. This study contributes to a transparent, and data-driven framework for agile story point estimation, connecting the gap between human expertise and Artificial intelligence decisions.

ACKNOWLEDGEMENT

This research would not have been possible to complete without the help I received from so many parties and people. First, I need to thank to the Sri Lanka Institute of Information Technology for giving myself to the chance to work on this research project and for helping me to build up both my research and academic skills.

And, I also am truly thankful to my supervisor, Dr. Mahima Weerasinghe. His support, guidance, and honest feedback in the last year really helped me to enhance my work and stay on the right track to complete the research project.

I am also thankful to my colleagues at SLIIT, who always shared useful advice and support to complete the research project on time. Their input and advice helped me to improve my thought process.

Finally, I owe the biggest thanks to my dearest wife and family. They were always behind me, motivating me, supporting me when I felt tired and always encouraging me to keep moving forward and keep on track. Their continuous support is the reason this research journey is possible.

TABLE OF CONTENTS

DECLARATION	2
ABSTRACT.....	3
ACKNOWLEDGEMENT.....	4
TABLE OF CONTENTS.....	5
List of Figures	8
List of Tables	8
Chapter 1 Introduction	9
1.1 Background: The Challenges and Role of the Story Point Estimations in Agile Development	9
1.2 Problem Statement.....	10
1.3 Research GAP	11
1.4 Research Questions	12
1.5 Research Objectives and Contributions.....	13
1.6 Significance of Study	14
1.7 Scope of the Study	15
Chapter 2 Literature Review	16
2.1 Introduction to Agile Estimation.....	16
2.2 Machine Learning Approaches for Story Point Estimation	17
2.2.1 Traditional Machine Learning Approaches	17
2.2.2 Deep Learning Approaches	18
2.2.3 Transformer-Based Approaches	18
2.2.4 Summary	19
2.3 Transformer-Based NLP Models for Story Point Estimation.....	19
2.3.1 The BERT Model in Effort Estimation.....	20
2.3.2 DistilBERT: Efficiency and Compactness	20
2.3.3 RoBERTa: Robust Optimization for Pretraining	21
2.3.4 DistilRoBERTa: Balancing Robustness and Efficiency.....	21
2.3.5 Relevance to Research Questions.....	22
2.3.6 Summary	22
2.4 Explainable AI in Software Engineering	22
2.4.1 Importance of Explainability in Agile Estimation	23
2.4.2 Local and Global Explainability	23
2.4.3 SHAP and LIME for Model Interpretability	24
2.4.4 Counterfactual Explanations and Captum	24
2.4.5 Explainability Evaluation Metrics.....	25

2.4.6 Summary	25
2.5 Human vs AI in Agile Estimation	25
2.5.1 Limitations of Human Estimation.....	26
2.5.2 Emergence of AI-Based Estimation	26
2.5.3 Evaluating Human–AI Performance Differences	27
2.5.4 Human Trust and Adoption of AI Estimators	27
2.5.5 Summary	28
2.6 Literature Review Summary.....	28
Chapter 3 Research Methodology	30
3.1 Research Design: A Quantitative Experimental Approach.....	30
3.2 Dataset and Preprocessing	30
3.3 Model Development and Fine-Tuning	41
3.4 Explainability Integration and Evaluation	46
3.4.1 SHAP Explanation	46
3.4.2 LIME Explanations	47
3.4.3 Counterfactual Explanation	49
3.4.4 Captum explanation.....	49
Chapter 4 Results	51
4.1 BERT-base Performance.....	51
4.2 DistilBERT performance	53
4.3 RoBERTa-base Performance	56
4.4 DistilRoBERTa-base Performance	59
4.5 Conclusion of Results	61
Chapter 5 Discussion.....	64
5.1 Introduction to the discussion	64
5.2 Model Performance and Comparative Analysis (RQ1)	65
5.3 Impact of Preprocessing and Data Augmentation (RQ2).....	66
5.3.1 Baseline vs Preprocessed Data	67
5.3.2 Effect of Adaptive Fibonacci Mapping	67
5.3.3 Semantic Data Augmentation Using T5	67
5.3.4 Context Injection for Semantic Enrichment.....	68
5.3.5 Integrated Pipeline Performance.....	68
5.3.6 Summary	68
5.4 Explainability and Human Alignment (RQ3).....	69
5.4.1 Dataset and Evaluation Design	69
5.4.2 Keyword Alignment and Numerical Overlap Analysis	70

5.4.3 Stability and Consistency of Explanations.....	70
5.4.4 Human Alignment and Expert Validation	71
5.4.5 Comparative Interpretation and Theoretical Reflection	71
5.4.6 Summary of Findings.....	71
5.5 Comparison Between Human and AI Estimation (RQ4)	72
5.5.1 Data and Evaluation Framework.....	72
5.5.2 Quantitative Comparison of Models and Human Estimations	73
5.5.3 Error and Deviation Analysis	74
5.5.4 Qualitative Comparison: Human vs AI Reasoning.....	75
5.5.5 Comparative Interpretation and Implications	75
5.5.6 Summary	75
5.6 Theoretical and Practical Implications	76
5.6.1 Theoretical Implications.....	76
5.6.2 Practical Implications	77
5.7 Limitations of the Study	78
5.7.1 Dataset Limitations	78
5.7.2 Experimental and Computational Constraints.....	78
5.7.3 Human Evaluation and Explainability Constraints	79
5.7.4 Methodological Scope	79
5.8 Future Work	79
5.8.1 Expanding Dataset Diversity and Real-World Validation.....	80
5.8.2 Advanced Model Architectures and Optimization.....	80
5.8.3 Integration of Inherently Interpretable Models	80
5.8.4 Human–AI Collaboration and Interactive XAI Tools	80
5.8.5 Broader Research Directions	81
5.9 Summary of Discussion	81
Chapter 6 Conclusion	83
6.1 Overview	83
6.2 Summary of Key Findings.....	83
6.2.1 Transformer Model Performance	83
6.2.2 Effect of Preprocessing and Data Augmentation.....	84
6.2.3 Explainability and Human Alignment.....	84
6.2.4 Human vs AI Estimation	85
6.3 Overall Contribution	85
6.4 Implications.....	86
6.5 Final Remarks	86

References	87
------------------	----

List of Figures

Figure 3.1 Distribution of Story Points across the Tawos Dataset	33
Figure 3.2 Story Point Data Box Plot	36
Figure 3.3 SHAP Explanations Visualizations	47
Figure 3.4 LIME Explanation Visualizations	48
Figure 3.5 Captum Explanations	50

List of Tables

Table 3.1 Tables in TAWOS dataset	31
Table 3.2 TAWOS Dataset Issue Table Description	32
Table 3.3 Story Points Wise Data Records Count	37
Table 3.4 Story Point to Label Mapping	38
Table 3.5 Label Wise Data Count	40
Table 3.6 Label Wise Data Record Count After Augmentation	41
Table 4.1 BERT-base Performance - Early Stopping Patience 10	51
Table 4.2 BERT-base Performance - Early Stopping Patience 4	52
Table 4.3 BERT -base Performance Comparison	53
Table 4.4 DistilBERT Performance - Early Stopping Patience 10	54
Table 4.5 DistilBERT Performance - Early Stopping Patience 4	54
Table 4.6 DistilBERT Performance Comparison	56
Table 4.7 RoBERTa-base Performance - Early Stopping Patience 10	56
Table 4.8 RoBERTa-base Performance - Early Stopping Patience 4	57
Table 4.9 RoBERTa-base Performance Comparison	58
Table 4.10 DistilRoBERTa-base Performance - Early Stopping Patience 10	59
Table 4.11 DistilRoBERTa-base Performance - Early Stopping Patience 4	60
Table 4.12 DistilRoBERTa-base Performance Comparison	61
Table 5.1 Quantitative XAI Measures (Expert selected and Model selected)	70
Table 5.2 Quantitative Comparisons of Model vs Human Expert Estimation	73