CHAPTER V
CONCLUSIONS AND RECOMMENDATIONS

This chapter concludes the study by presenting its principal findings and recommendations for future research. The conclusions summarize the main findings and address the research objectives established at the beginning of this study. The recommendations provide directions for future work, including both methodological improvements and opportunities for practical implementation. Together, these sections summarize the contributions of this research while identifying potential avenues for further investigation.

5.1 Conclusions

Based on the results and discussion presented throughout this study, the following conclusions can be drawn:

1. The proposed stacking ensemble model, which combines XGBoost and LightGBM as base learners with Logistic Regression as the meta-learner, demonstrated more consistent performance than the individual models across the evaluated distribution shift scenarios, including both unseen attack and unseen endpoint settings. Although LightGBM most frequently achieved the highest Recall and Macro-F1 scores, the stacking ensemble never ranked last in any evaluated combination of metrics and scenarios and achieved an average ranking of 1.67, compared with 1.57 for LightGBM. In addition, the stacking ensemble most frequently achieved the highest PR-AUC. The error analysis further showed that the effects of distribution shift varied across attack types and endpoints, indicating that model robustness depends not only on the presence of distribution shift but also on the characteristics of the underlying data. Overall, the experimental results suggest that the stacking ensemble provides greater robustness to distribution shift than either XGBoost or LightGBM when robustness is evaluated in terms of performance consistency across diverse testing conditions.

2. The streaming inference evaluation demonstrated that the relative performance observed in the offline evaluation was largely preserved under continuous inference. Across most evaluation scenarios, the model rankings obtained during streaming inference remained consistent with those observed offline. LightGBM most frequently achieved the highest Recall, whereas the stacking ensemble most frequently achieved the highest PR-AUC and never ranked last. Overall, the stacking ensemble obtained the best average ranking during the streaming evaluation (1.52), outperforming LightGBM (2.05), which indicates greater performance consistency across the evaluated streaming scenarios. This improved consistency, however, came at the cost of higher computational overhead. The stacking ensemble consistently exhibited the highest P95 latency and the lowest throughput, whereas LightGBM achieved the lowest P95 latency and the highest throughput. These findings demonstrate a trade-off between predictive robustness and computational efficiency, suggesting that the choice of model should depend on the deployment requirements of the intended intrusion detection system.

5.2 Recommendations

Based on the findings of this study, the following recommendations are proposed for future research:

1. Future studies should evaluate the proposed approach using additional intrusion detection datasets or perform cross-dataset evaluations to assess the generalizability of the models under network environments exhibiting more diverse traffic characteristics.
2. The distribution shift analysis in this study focused primarily on several key features selected using Permutation Feature Importance (PFI). Future research could investigate more comprehensive approaches to distribution analysis by incorporating a larger set of features or employing more advanced feature representation and visualization techniques.
3. Future work could explore alternative stacking ensemble configurations by investigating different combinations of base learners and meta-learners to

better understand their impact on classification performance, robustness to distribution shift, and computational efficiency.

4. In addition to stacking, future studies could evaluate other ensemble learning approaches, such as blending, voting ensembles, or more recent boosting-based methods, to compare their generalization capability and robustness under various forms of distribution shift.
5. The streaming inference evaluation conducted in this study was performed in a simulated environment. Future research should validate the proposed models in operational network environments to assess their performance under continuously evolving traffic distributions and long-term deployment conditions.