CHAPTER 8

STATISTICAL VALIDATION AND PROPOSED MODEL COMPARISON IN DDOS ATTACK DETECTION

8.1 Introduction

DDoS attacks are becoming more sophisticated and frequent, detecting them is a crucial cybersecurity concern. Accurate detection mechanisms need reliable statistical validation techniques for enhancing the detection models' reliability (Ibrahim, Z.K. and Thanon, M.Y., 2021). This chapter discusses ANOVA, PTest and FTest for statistical analysis and then evaluates performance of several proposed models that depends on ML and DL methods for detecting DDoS attacks. It proposes a number of innovative models and methods for DDoS detection.

In particular, when using ANOVA for comparing the accuracy of different machine learning models, between groups and within groups variances are critical. Within group means Coefficient of variance measures across groups while between groups variance focuses on the variance in the mean accuracy score of different models. This variance is used to identify whether or not the use of a particular model has a bearing on the performance being achieved. On the other hand, within the group variance will estimate how much the static accuracy will vary within the same group in the model due to uneven distributions across iterations or splits. This captures the variability inherent in the model evaluation process arising from issues such as random initialization of the network or partitioning of the data. Specifically, ANOVA tests compare these variances to determine if the differences in model accuracy that we observe in our evaluation are statistically significant or if they might have been caused by chance. If the "between groups" variance is considerably higher than the "within groups" variance, it indicates that the choice of model substantially affects accuracy (Kim, T.K., 2017) (Rouder J.N et al., 2016).

Other metrics include SS, df , MS , F-value, and p-value. To test the hypothesis that group averages are equal using the F-value, which is proportion between the mean square among groups to the mean square inside groups. A greater probability of different group means is indicated by a higher F-value. Assuming the null hypothesis is true, the p-value evaluates the statistical significance of the F-score and shows the likelihood of finding an F-value as severe as the computed one. Stronger evidence against the null hypothesis is

indicated by a decreased p-value. Major differences between the group means are shown when the p-value is lower than the traditional significance threshold, which is often 0.05, and the null hypothesis is neglected. In this process of statistical validation, hypotheses are essential. The null hypothesis (H) in this paradigm assumes that the accuracy of the compared models does not vary significantly. However, the alternative theory (H) proposes that there is at least one model has significantly difference in accuracy. By testing these hypotheses, it can be can determined that whether the observed differences in model performance are statistically eloquent or simply the result of random variations.

Phase I: Ensemble-Based CFFS with DT Classifier

The proposed models include the Combined Filter for Feature Selection (CFFS) method coupled with Decision Tree (DT) classifiers, which showed significant gains in execution speed and accuracy when compared to other techniques assessed in Phase I. According to the ANOVA results, the suggested method and the other compared methods differ statistically significantly. The ANOVA table 8.1 and figure 8.1 demonstrate a significant F-value of 9.34 with a very low p-value indicating 0.0068, where the default p-value must be 0.0001.

Table 8.1 ANOVA Results for Proposed CFFS-DT method and other methods evaluated in Phase I

Source of Variation	SS (Sum of Squares)	df (Degrees of Freedom)	MS (Mean Square)	F-value	p-value
Between Groups	2752.308	9	2752.308	9.34	0.0068
Within Groups	5306.0245	18	294.7791		
Total	8058.3325	19			

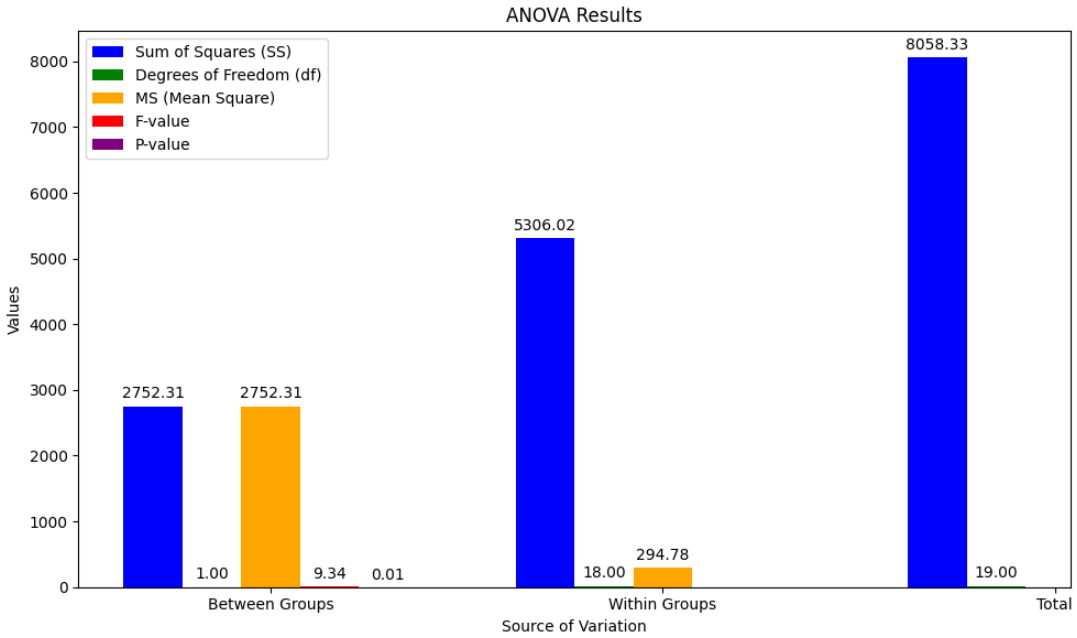

Figure 8.1 Bar chart on ANOVA Results for Proposed CFFS-DT method and other methods evaluated in Phase I

The figure 8.2 shows an F-distribution curve with df1 = 1 and df2 = 18, used in hypothesis testing. The Probability Density is shown by the blue curve, and the red-shaded area marks the critical region for a 0.05 significance level. The red dashed line at an F-value of 4.41 is the critical value. The green dashed line at an F-value of 9.34 represents the calculated F-statistic.

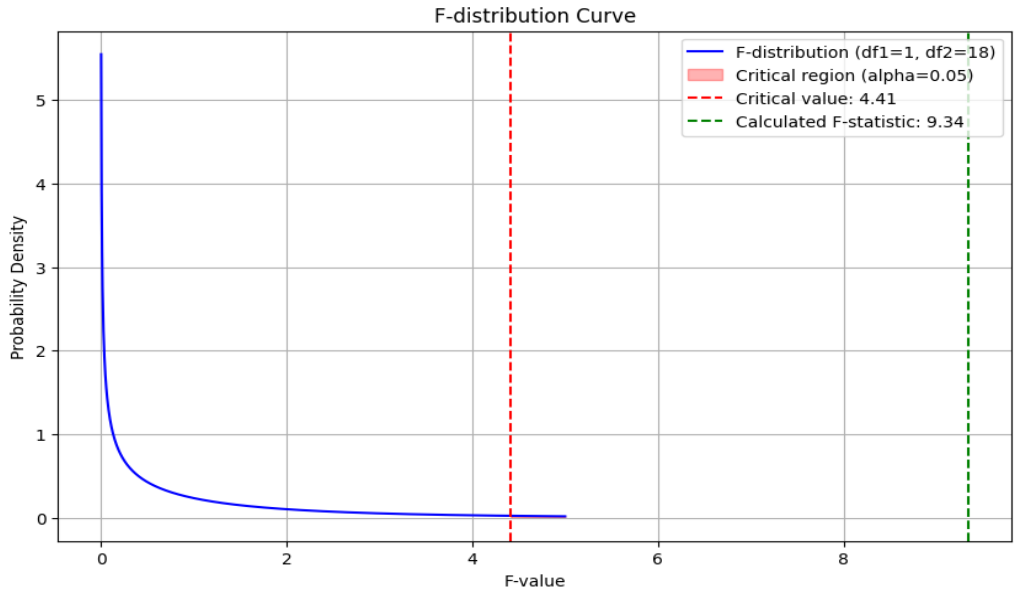

Figure 8.2. F-Distribution Curve with Critical and Calculated F-Statistics for Hypothesis Testing on proposed approach in Phase I

Phase II: Hybrid Feature Engineering with Improved Dragonfly Optimization Algorithm (IDOA) and Decision Tree (DT)

The proposed hybrid feature engineering model integrates the Improved Dragonfly Optimization Algorithm (IDOA) with Decision Tree (DT) classifiers to improve multi-vector DDoS flooding attack detection. This approach demonstrates a marked improvement in accuracy and execution time over other methods evaluated in Phase II. The ANOVA results provide strong validation for this approach. The ANOVA table (Table 8.2) and corresponding bar graph (Figure 8.3) illustrate a significant F-value of 22.02 with a very low the difference is mathematically relevant, as shown by the p-value of 0.00029.

Table 8.2 ANOVA Results for Proposed IDOA-DT method and other methods evaluated in Phase II

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Value	p-Value
Between Groups	1050.6006	5	1050.601	22.0192	0.00029
Within Groups	715.6929	15	47.7129		
Total	1766.2934	17			

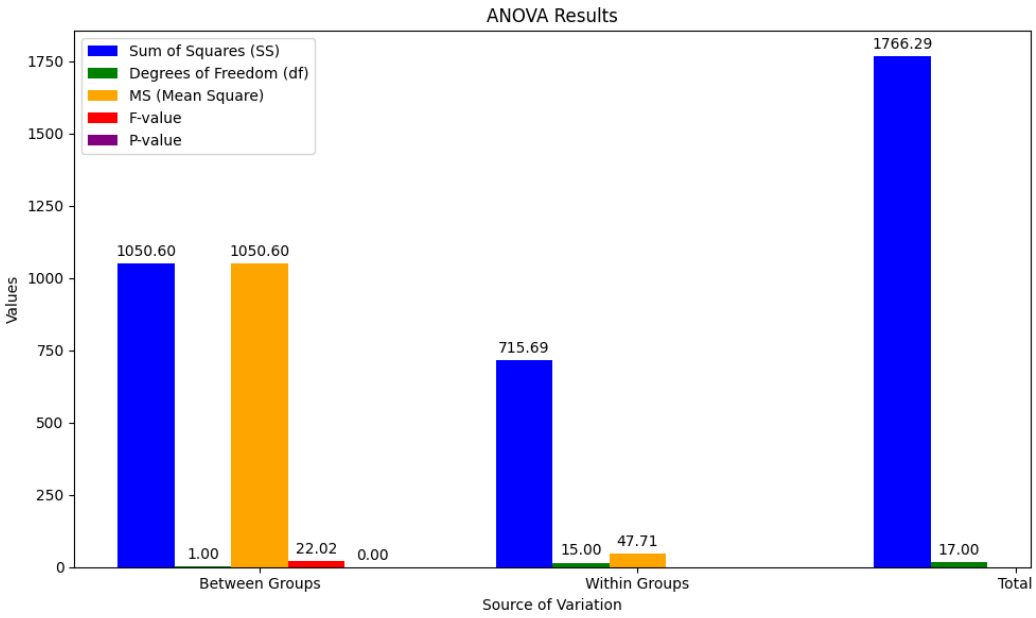

Figure 8.3 Bar chart on ANOVA Results for Proposed IDOA-DT method and other methods evaluated in Phase II

The F-distribution curve (Figure 8.4) highlights the critical region (alpha = 0.05) and the calculated F-statistic. The calculated F-statistic of 22.02 significantly exceeds the critical value of 4.54, further confirming the proposed hybrid IDOA-DT method's efficacy in contrast to other Phase II approaches that were assessed. The model efficiently chooses the most relevant features by using the Improved Dragonfly Optimization Algorithm for feature engineering and selection. This enhances the Decision Tree classifier's ability to identify multi-vector DDoS assaults.

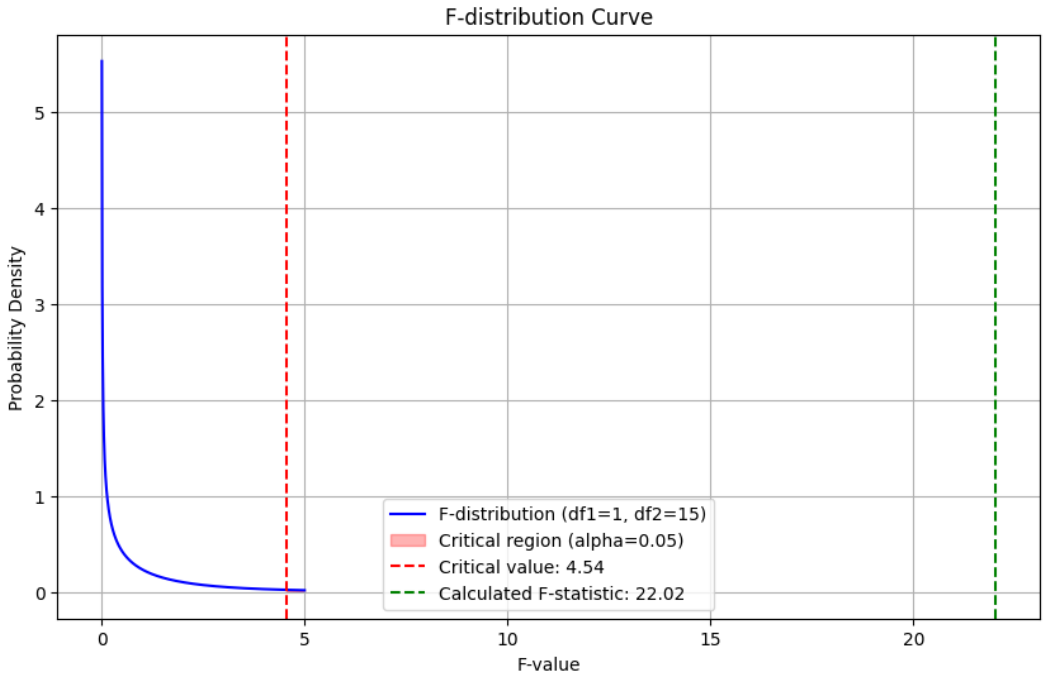

Figure 8.4 F-Distribution Curve with Critical and Calculated F-Statistics for Hypothesis Testing on proposed approach in Phase II

Phase III: Panthera Leo Optimized Multilayer Feed Forward Learning for Multi Vector DDoS Attack Detection

This model employs Panthera Leo Optimization technique to increase multilayer feedforward networks' accuracy and learning efficiency. The model's effectiveness was validated through comprehensive statistical tests, showing significant. Based on the various performance metrics of different classification approaches across various dataset splits the ANOVA test has been carried out. The ANOVA results indicate statistically significant differences between the various detecting algorithms' performance metrics. The statistical significance of the group mean differences is confirmed by the low p-value (0.01445) and big F-value (12.0114).

Comprehensive statistical validation through ANOVA confirmed significant improvements in performance metrics. The ANOVA results (Table 8.3 and Figure 8.5) indicate statistically significant differences between the classification approaches, with the Panthera Leo model demonstrating noticeable performance across various dataset splits (80:20, 70:30, 60:40, 50:50). The efficacy of the suggested model in identifying multi-vector DDoS assaults is shown by the consistently higher values it obtains. This strategy is well validated by the ANOVA findings.

Table 8.3 ANOVA Results for Proposed PLO-MLFFN method and other methods evaluated in Phase III

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Value	p-value
Between Groups	1292.9247	4	323.231	12.0114	0.01445
Within Groups	546.1173	15	36.4078		
Total	1839.042	19			

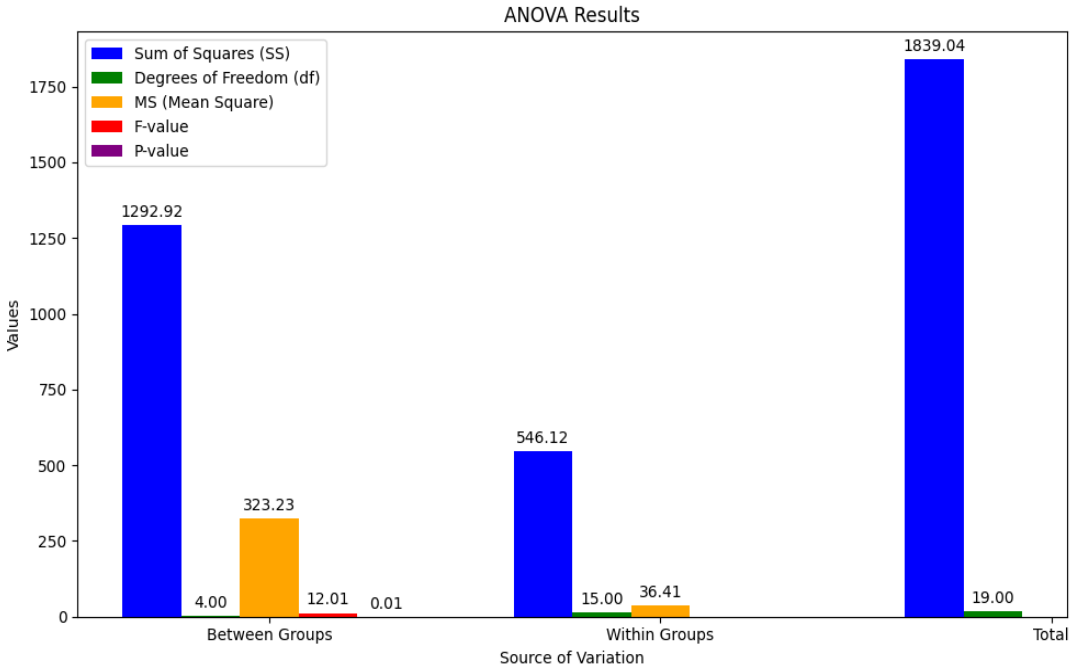

Figure 8.5 Bar chart on ANOVA Results for Proposed PLO-MLFFN method and other methods evaluated in Phase III

The F-distribution curve (Figure 8.6) highlights the critical region (alpha = 0.05) and the calculated F-statistic. The calculated F-statistic of 12.0114 significantly exceeds the critical value of 3.06, further confirming the efficacy of the proposed PLO-MLFFN approach in contrast to other techniques assessed during Phase III.

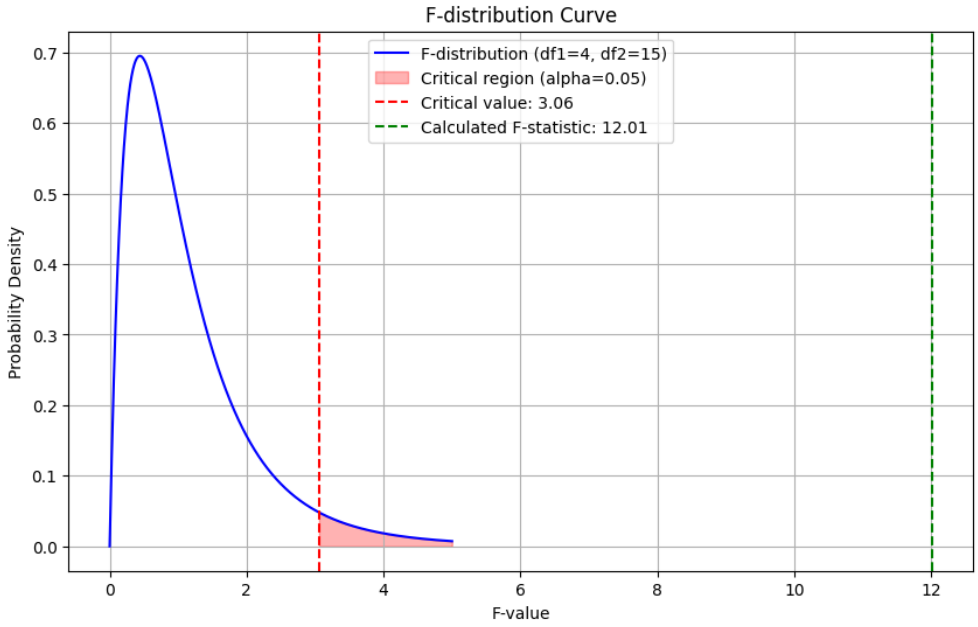

Figure 8.6 F-Distribution Curve with Critical and Calculated F-Statistics for Hypothesis Testing on proposed approach in Phase III

Phase IV: Attention Enabled Gated Recurrent Network (AEGRN) and Deep Feed Forward Networks

The ensemble model of Attention Enabled Gated Recurrent Network (AEGRN) and Deep Feed Forward Networks (DFFN) introduced in phase IV demonstrates superior performance in detecting DDoS attacks across multiple datasets. This model showed high accuracy, and better performance. The effectiveness F-test and ANOVA were utilized to validate the suggested model, highlighting significant improvements over methods.

The ANOVA results, presented in the table 8.4 and figure 8.7, gives a statistical validation of the functional metrics for different models. The ANOVA analysis confirms statistically significant differences between the performance of the different detection models, showing that the suggested AEGRN and DFFN ensemble model performs noticeably better than other techniques in the identification of DDoS attacks on a variety of

datasets. This validation highlights how reliable and effective the suggested model is in actual DDoS detection situations.

Table 8.4. ANOVA Results for Proposed AEGRN-DFFN method and other methods evaluated in Phase IV

Source of Variation	SS (Sum of Squares)	df (Degrees of Freedom)	MS (Mean Square)	F-value	p-value
Between Groups	1524.86	3	508.29	12.63	0.0002
Within Groups	1610.52	40	40.26		
Total	3135.38	43			

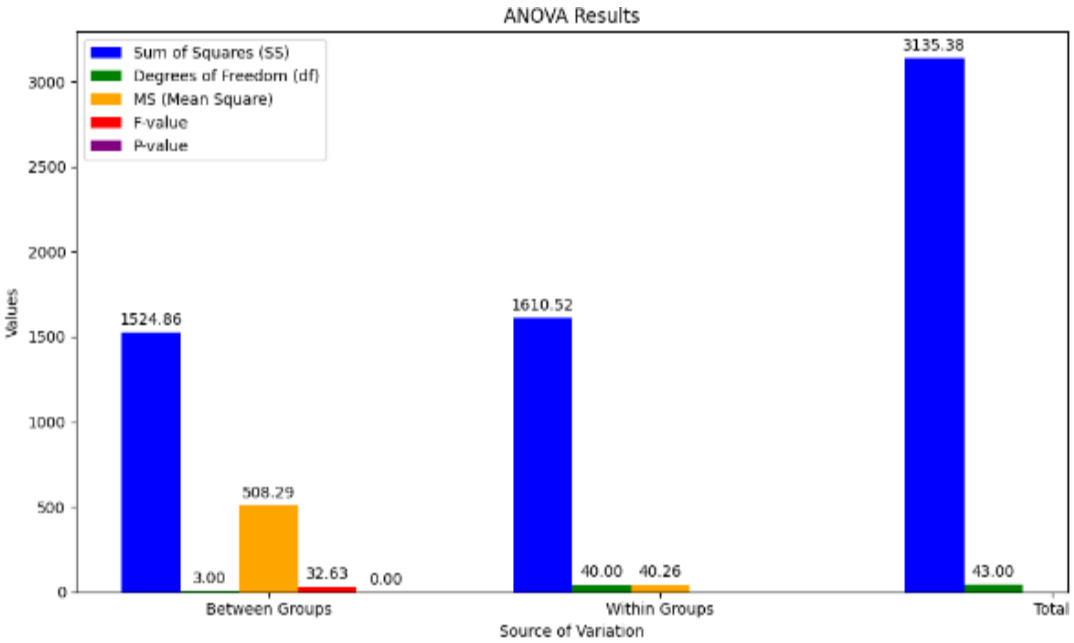

Figure 8.7 Bar chart on ANOVA Results for Proposed AEGRN-DFFN method and other methods evaluated in Phase IV

The figure 8.8 shows an F-distribution curve with df1 = 3 and df2 = 40, used in hypothesis testing. The probability density is shown by the blue curve, and the crucial region with a significance level of 0.05 is shown by the red-shaded area. The crucial value is shown by the red dashed line with an F-value of 2.84. The computed F-statistic is shown by the green dashed line with an F-value of 12.63. The null hypothesis is neglected because the computed F-statistic is higher than the crucial value, demonstrating a substantial difference in group variances.

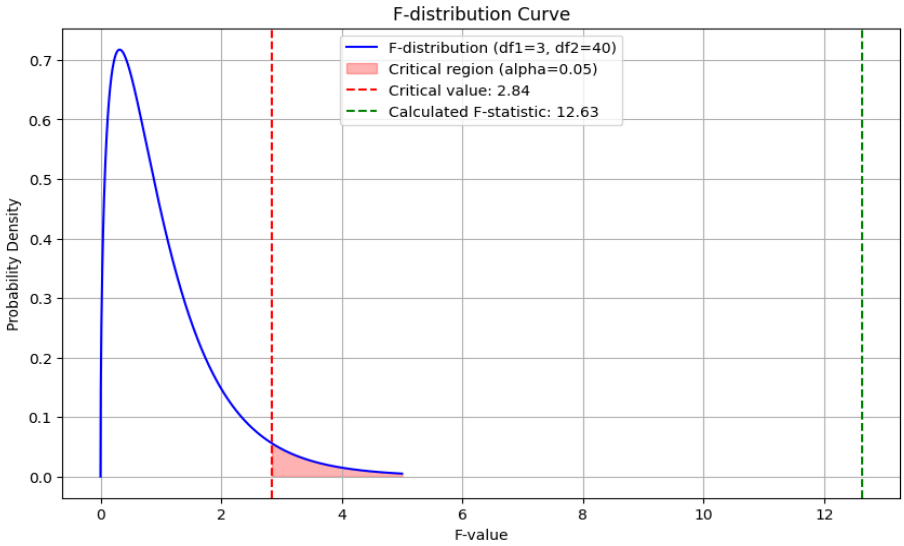

Figure 8.8 F-Distribution Curve with Critical and Calculated F-Statistics for Hypothesis Testing on proposed approach in Phase IV

Having a p-value around 0.0001 to 0.0002 and an F-value between 9 and 22 is generally considered very strong evidence of statistical significance, particularly for the model of an IDS. A p-value in this range indicates a very low likelihood that the observed findings are not the product of random chance since it is far lower than typical significance values (such as 0.05 or 0.01). The null hypothesis is strongly refuted by this, showing that the model's performance improvements are statistically significant. The ANOVA test is performed to compare the efficacy of four distinct statistical models: CFFS-DT, IDOA-DT, PLO-MLFFN, and AEGRN-DFFN. The results were visually seen in Figure 8.9's bar chart, indicating the F and P values for each model.

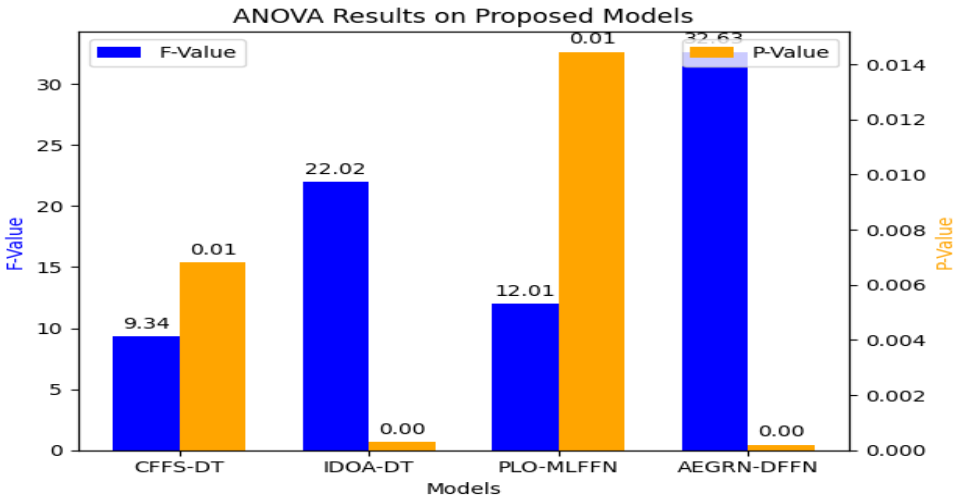

Figure 8.9 ANOVA results comparison for the proposed models

The AEGRN-DFFN model exhibited the greatest F value of approximately 32.63, suggesting a significant difference among groups as opposed to within groupings, but keeping a low P value of around 0.0002, which is indicative of strong evidence against the null hypothesis. Similarly, the other models also demonstrated low P values, implying statistical significance. However, the AEGRN-DFFN model stood out as potentially the most statistically significant model, given its higher F value. These results are still preliminary and need to be investigated more, using raw data and the analysis method which was employed. However, given that all the P values from all the models are less than 0.05, the models might indeed have importance in their own areas of use. In the presented study, both the high statistical significance, given by the low p-value, and the high F-value support the conclusion that the obtained results are not caused by random fluctuations and are actually representative of the performance of the model, which can, therefore, be considered a reliable tool to detect intrusions. These advancements help in creating sophisticated intrusion detection technologies that will assist in shielding computer-based system architectures from the ever emerging DDoS attacks.