## **Two Major Findings**

### **1. High False Positive Rate (FPR) in Existing Models**

* **Problem**: Many anomaly detection methods trigger too many false alarms. In critical systems, even a single false alert may cause wasted resources, panic, or incorrect mitigation measures.
* **Research Insight**: The focus must shift from overall accuracy (which can be misleading in imbalanced datasets) to metrics like **precision, recall, and especially false positive rate**.
* **Research Gap**: Existing works achieve **99–100% accuracy/F1 scores**, but these often hide **overfitting issues** or weak generalization to real-world traffic.

### **2. Dataset Preparation Problems**

* **High Dimensional Data**: CICIDS-2017/2018 has ~80 features; feature extraction/selection is difficult but necessary for efficiency and generalization.
* **Class Imbalance**: Many attack classes are underrepresented, making it hard for models to learn. **SMOTE (Synthetic Minority Oversampling Technique)** or other resampling methods are frequently used.
* **Overfitting**: Many papers report “too good to be true” results (99–100% scores), likely due to poor validation or leakage between train/test sets. Regularization and robust validation are needed.
* **Missing Values**: Example — 2,867 missing rows in CICIDS, which must be handled.
* **Normalization**: Needed for ML/DL models since raw features have very different scales (e.g., bytes vs. packet counts).
* **Efficiency Constraint**: Time and computational power efficiency matter, so **dimensionality reduction (feature selection)** is critical.

## **Research Timeline (Bi-Weekly with Professor Meetings)**

You’ve already completed **Week 1–2: Literature Review**. Below is a proposed plan for ~14 weeks (flexible to exams).

### **Weeks 1–2 (Already Done)**

* **Task**: Literature review on anomaly detection using CICIDS and related datasets.
* **Output**: Major findings (false positives + dataset preparation problems).

### **Weeks 3–4: Data Preparation – Initial Experiments**

* Handle **missing values** (drop rows vs. impute).
* Apply **normalization/scaling** (MinMax, StandardScaler).
* Analyze **class imbalance** → test SMOTE and compare with undersampling.
* Start **feature selection/reduction** (e.g., PCA, correlation-based, feature importance via tree models).
* **Meeting Goal**: Present dataset cleaning pipeline and first feature-selection results.

### **Weeks 5–6: Exam Break + Light Work**

* Focus shifts to **academic exams** (only minimal experiments).
* Parallel: write **short notes/documentation** on preprocessing methods tried so far.

### **Weeks 7–8: Baseline Models**

* Train baseline models (Logistic Regression, Random Forest, LightGBM, simple ANN).
* Evaluate with **precision, recall, F1, ROC, and especially FPR**.
* Compare models trained on **full vs. reduced features**.
* **Meeting Goal**: Present baseline results and identify which models/feature sets look promising.

### **Weeks 9–10: Advanced Models & Regularization**

* Experiment with **regularized deep learning models** (Dropout, L1/L2).
* Compare traditional ML vs. DL performance.
* Investigate **time-series nature** of traffic (RNN/LSTM/GRU vs. tabular ML).
* **Meeting Goal**: Show whether DL improves results without overfitting.

### **Weeks 11–12: False Positive Minimization**

* Apply **threshold tuning**, cost-sensitive learning, or anomaly scoring.
* Try **ensemble approaches** (stacking, boosting, hybrid ML-DL).
* Compare performance focusing on **minimizing FPR**.
* **Meeting Goal**: Report improvements in FPR vs. baseline.

### **Weeks 13–14: Final Evaluation + Paper Writing**

* Perform **cross-validation** and confirm results.
* Prepare visualizations (ROC curves, confusion matrices, feature importance).
* Start **research paper draft**:  
  + Introduction & Related Work (from literature review).
  + Methodology (from Weeks 3–10).
  + Results & Discussion (Weeks 11–12).
* **Meeting Goal**: Get professor’s feedback on paper draft.

By the end of ~14 weeks, you’ll have:

1. A well-documented **preprocessing + feature selection pipeline**.
2. Models tested and evaluated with **FPR focus**.
3. A **research paper draft** ready for refinement and submission.