*An abstract on*

**Anomaly Detection - Based on over-sampling PCA**

*Submitted in partial fulfillment of the requirements*

*For the award of the degree of*

**BACHELOR OF TECHNOLOGY**

*in*

Computer Science & Engineering

*By*

**M.Dhenuka Datta [164G1A0521]**

**B.Divya Bharathi [164G1A0523]**

**G.Kaleem [164G1A0534]**

**P.Chakrapani [164G1A0517]**

Under the Guidance of

Ms. Karamala. Hema Latha M.Tech.,

Assistant Professor



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY: ANANTAPUR**

**(Accredited by NAAC with ‘A’ Grade, Affiliated to JNTUA, Approved by AICTE, New Delhi)**

**2019-2020**

# ABSTRACT

Anomaly detection is the process of identifying unexpected items or events in data sets. It is often applied on unlabeled data which is known as unsupervised anomaly detection. It has been an important topic in machine learning and data mining. In data mining , anomalies are referred to as outliers , noise , deviations and exceptions. It is used in real-world applications which includes intrusion or credit card fraud series detection .

In existing system , the process of detecting anomalies is done in batch mode by using traditional techniques like Distributed or statistical, Density based, Distance based methods. Principal Component Analysis is a modern method. It is a well known unsupervised dimension reduction method which cannot be applied to real-world large-scale applications. For example, in power method for producing the approximated PCA solutions we need to store the data of covariance matrix which leads to high computation cost, memory requirements and cannot be easily extended to applications with streaming data.

In our proposed system, we are using online over-sampling Principal Component Analysis (osPCA) in order to overcome the above problems. We are using an online updating technique in our osPCA so that there is no need to store the data of covariance matrix for producing the approximated PCA solutions , which in turn reduces the computational cost and memory requirements for large data sets.

Date: 23 December 2019 Signature of the Guide