Abstract

Load balancing (LB) is the process of distributing the workload fairly across the servers within the cloud environment or within distributed computing resources. Workload includes processor load, network traffic and storage burden. LB's main goal is to spread the computational burden across the cloud servers to ensure optimal utilization of the server resources. Cloud computing (CC) is a rapidly growing field of computing that provides computing resources as a product over the internet. This paper focuses on the issues within Cloud Load Balancing (LB) that have attracted research interest. The paper also mainly focused on uncovering machine learning models used in LB techniques. The most common algorithms in the reviewed papers included Linear Regression, Random Forest classifier (RF) Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long-Short Term Memory- Recurrent Neural Network (LSTM -RNN). The criteria for LB technique was identified through performance metrics like throughput, response time, migration time, fault tolerance and power saving. The paper adjourns by identifying research gaps found in the reviewed literature.