I. IMPORTANT FINDINGS

In this section, we will discuss the important findings that were made by doing the literature survey.

- Load prediction over a specific period of time comes under time series analysis in ML
- There are many ways to perform time-series predictions
- AR model, MA model, ARMA model, ARIMA model are some of the most popular approaches for time-series predictions and among them, the autoregressive integrated moving average model is the most accurate
- Other than that Neural networks can be used to perform time-series predictions'
- LSTMs or Long short term is a special kind of recurrent neural network where information persistence can be achieved. Other than that there is a feedback connection to improve learning that differentiates it from a regular feedforward network.
- Bidirectional LSTM neural network is a special extension of the LSTMs the input vectors are analyzed in both the forward direction and the backward direction by 2 independent LSTMs.
- All the NNs try to minimize the cost function using gradient descent.
- Bi-LSTM is mainly used for time series data predictions.
- The hidden layers transform the input vector into the output vector.
- From the studies, it is found that as the number of steps in the NN increases, the accuracy and the efficiency of Bi-LSTM increases compare to the ARIMA and LSTM model
- K-shape algorithm is a novel clustering algorithm that is used to identify the most important features to the input vector of an LSTM

II. IMPROVEMENT

The existing proposed solutions only consider the metrics either at an orchestrator level or a cluster level to make time-series predictions for the workload. But in essence, a combination of them plays a role in the performance hindrance of the system. Therefore it is important that a method that collectively takes into account all the metrics at all abstraction levels of virtualization is taken into account before predictions are made.

As discussed under [2] an MSM component can be extremely important to any prediction system which uses categorization algorithms such as k-shape to identify the most important features for the prediction algorithm. This logic can be extended further to identify the training features not only in container level metrics but also in metrics gathered at a cluster level to create the input vector.

Further, a proactive mechanism for the auto scaler to dynamically scale up the clusters can be proposed since the bidirectional LSTM uses a width of w and identifies past trends and future tend to predict the values which helps in accurately predicting the load throughout the operation time.

Also, it was observed that in some scenarios, as the number of steps is lower, the ARIMA-based models yield high accuracy than LSTM based approaches. Therefore, a mechanism can be implemented to predict the workload using a combination of the models such that the accuracy is always high in all circumstances.