SUMMARY

The initial thesis objectives were to investigate current methods of HPO and analyze from the simplest to the state-of-the-art, in order to develop an implementation to the University of Tartu's HPO R package, SmartML package, whose primary objective was that it would, at the very least, provide the same amount of performance as the current implementation. By analyzing automated machine learning literature it was possible to see that research in HPO is split into two points of focus, and the first one is more concentrated on getting machine learning algorithms' hyperparameter configurations to be tested as fast and as wide as possible. The second one hovers over finding the best configuration search space as early as feasible, assuming that the distribution might have either a single global optima or very few considerably clustered together.

The first focus was developed from the random search standpoint, in which hyperparameter configurations were to be sampled uniformly, effectively allowing the scientist to have an overview of multiple points in the configuration space, however, this was proven to be an incomplete approach, since some machine learning algorithms can be very slow, meaning that an approach with such breadth would take too long. The solution to this is to explore the concept of budget, or resource, which is to define a variable that can significantly impact the running time of a machine learning algorithm. In the thesis work, the budget concept to be used was of dataset size, since machine learning algorithms can have very poor exponential scalability and thus reducing its size can lead to many more hyperparameter configurations being tested, as it was shown on chapter 3, in which successive halving, which essentially is simply a random search that is budget-aware, it merely ranks and stops giving more data to models that are underperforming, managed to test almost 18 times as many configurations as plain random search did in the same time.

Even though successive halving performed very well, it still didn't perform optimally in situations where the machine learning algorithm to be tuned depended on a large amount of data. As it was shown, support vector machines are very sensitive to the dataset size and can even fail to predict in the early stages of successive halving, which trains the algorithms with very little data, thus there needed to be some kind of automated successive halving parameter tuning, in order to increase the number of datapoints fed to models at an earlier stage, and reduce the number of models that are tested. A solution to this was found on the hyperband algorithm, which adaptively experiments with different amounts of budgets and models to be tested in order to check what is the most optimal strategy to tune the designated machine learning algorithm within a specified budget constraint.

Despite hyperband's clever approach to perfecting successive halving, it still suffered from the primordial issue of random search, which is that it is not aware of which exploration direction leads to the best performing configurations, that is, even if it found the global optima already it won't explore its neighbourhood and instead will keep uniformly exploring the entirety of the search space. Adding guidance to hyperband was proven to be highly effective, with very visual results, as shown on chapter 6, that adding kernel density estimators to model which configurations are performant and sub-par resulted in a very computationally fast, demonstrated empirically on chapter 5, way to add an automated self-directing mechanism to it. This algorithm is called BOHB, Bayesian Optimization Hyperband.

After analyzing, implementing and profiling the researched algorithms, it was possible to test them using real-life datasets that posed a varied amount of challenges. The final test consisted of, similarly to the real-life scenario in which HPO methods would be used, letting all algorithms run for a specified amount of time, as a budget, to reach the highest accuracy as possible. A short amount of time, 2 minutes, and a longer one, 10 minutes, was picked and hundreds of runs over these times were done and recorded. SmartML's current implementation was shown to not reach the highest accuracy within the given time in the majority of the tests, with it losing almost every single one of the tests during the 10 minutes run, effectively showing that the solutions that were implemented as the goals of this thesis met the designated expectations.

BOHB has many hyperparameters to tune, and thus it would be paradoxical if a hyperparameter tuning algorithm would be very dependent on the proper tuning of its hyperparameters itself, hence the robustness to changing its algorithms, assuming that they are independent, was tested. This robustness test was done in the final part of the last chapter, and it was done by comparing multiple BOHB hyperparameter configurations against itself and a hyperband run. The statistical test done was a t.test, with roughly 97 degrees of freedom, two-tailed and aiming to measure whether the difference in means from the hyperparameter configurations distribution was statistically significant. The finds showed that the vast majority of the hyperparameter configurations were not statistically significant from each other, only from the hyperband run, effectively, arguably, proving the claim made on its paper, that it indeed is highly invariant to its hyperparameters.

Concluding the main summary of research, it can be said that the results obtained successfully satisfied the given tasks. The first task, of implementing bayesian optimization, in this case TPE, and integrating it to a multi-armed bandit problem, the second task, using hyperband with TPE in HPO, were done over the course of chapters 3 to 5, while the last task of evaluating the algorithm and verifying its effectiveness was achieved on chapters 6 and first half of 7, while the last half of 7 was dedicated to investigating the hyperparameters of the optimization algorithm BOHB, which also indirectly proved that adding TPE to hyperband indeed yields statistically relevant different results.