**Title of Invention:** Method and System for Taxi Demand Prediction Using A Neural Network Model

## Technical Field

The present invention relates to a neural network model for taxi demand prediction, in particular a macro-micro multiview neural network model for taxi demand prediction.

## Background

In many large cities, taxis or ride-hailing services are playing an increasingly important role in providing point-to-point transportation services. However, a major problem in operating any type of point-to-point transportation service is the imbalances of vacant cars and passenger demands. As these imbalances are highly dynamic, and when drivers are left on their own to rebalance themselves based on historical information or their own limited local observations, the resulting system performance would be far away from optimal.

In existing taxi or ride-hailing demand predictions, a wide range of methods, from the classical statistical approaches to the more recent deep-learning-based techniques have been used. Both taxi and ride-hailing passenger demand predictions are well-studied in the literature, however, they are fundamentally different despite their similarities. A major differentiator is that while ride-hailing demands are all pre-booked (thus allowing operators to know exactly where the demands are, whether they are eventually satisfied or not), taxi demands can come from street-hail, taxi stands, or pre-booked channels. This makes taxi demand prediction more challenging.

In the taxi industry, existing systems are only capable of showing current demand hotspots to drivers. Even when the information is accurate, drivers may not find such information useful, as it provides only a snapshot of the current situation, and not prediction into the future. Furthermore, such demand hotspot visualization also lacks the capability in addressing demand-supply imbalances and does not provide guidance to individual taxi drivers.

Ride-hailing companies have a better form of guidance, as they can guide drivers indirectly by offering dynamic pricing (or surge-pricing). But unless the dynamic pricing scheme is personalized, similar supply-demand imbalance could still be happening (i.e., inexperienced drivers would be “chasing” demand hotspots or high surge prices, and effectively just shift congestion from one area to another area).

Existing methods are thus limited and may be further improved.

**Summary**

In a first aspect of the invention, there is provided a computer-implemented method for predicting taxi demand. The method comprises generating, by a three-dimensional spatiotemporal model, a current demand count for a plurality of unit grids based on a current taxi dataset comprising current demand and supply of a taxi fleet, wherein the three-dimensional spatiotemporal model comprises a convolutional neural network trained on a first dataset of demand and supply of taxis at the plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods.

Preferably, each three-dimensional image is a $K \times K \times h$ image, wherein h represents a number of historical time period, K represents a size of a neighbourhood proximate to the unit grid, and each pixel in the three-dimensional image has dimensions $(x, y, m)$, where $(x, y)$ refer to a coordinate of the unit grid and m refers to a specific time period. More preferably, data of each pixel is stored in three channels respectively encoding the number of trips originating in the coordinate of the unit grid, the number of trips ending in the coordinate of the unit grid, and the number of vacant taxis observed in the unit grid in the specific time period.

Preferably, the method further comprises generating, by a micro-movement model, a current micro-movement modifier, wherein the current micro-movement modifier reflects a current probability of taxi demand in a plurality of road links based on the current taxi dataset comprising current elapsed time of most recent vacant taxi for

the road links and current summary statistics of trips in the road links, wherein the micro-movement model is trained on a second dataset of elapsed time of most recent vacant taxi for the road links and summary statistics of trips in the road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and to output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region; and generating, by an integrating neural network, a current combined taxi demand prediction value based on the current demand count and the current micro-movement modifier, wherein the integrating neural network is trained to combine and assign weightage on the demand count and the micro-movement modifier. More preferably, the micro-movement model comprises at least two stacked Long Short-Term Memory (LSTM) networks.

Preferably, the method further comprises generating, by a hybrid LSTM-TCN model, a current exogenous taxi demand modifier based on both the current taxi dataset and additional dataset comprising current exogenous factors, wherein the hybrid LSTM-TCN model comprises at least two LSTM networks and a temporal convolutional network (TCN) trained on the first dataset and a third dataset of exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect demand and supply of taxis, wherein generating, by the integrating neural network, the current combined taxi demand prediction value includes generating, by the integrating neural network, the current combined taxi demand prediction value based on the current demand count, the current micro-movement modifier and the current exogenous taxi demand modifier, wherein the integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier. More preferably, the current taxi dataset and additional dataset are provided to the at least two LSTM networks and outputs from the LSTM networks serve as inputs to the TCN to output the current exogenous taxi demand modifier. As an example, the exogenous factors are selected from the following: meteorological conditions in the time period, temporal data in the time period and taxi-related data in the time period.

Preferably, the current taxi dataset comprises a current taxi location, and the method further comprises identifying at least one unit grid or road link with taxi demand proximal to the current taxi location based on the current demand count or the current combined taxi demand prediction value. More preferably, the method further comprises generating a personalised recommendation for an individual taxi driver based on the current demand count or the current combined taxi demand prediction value and location of vacant taxis.

In a second aspect of the invention, there is provided a non-transitory computer readable medium comprising instructions which, when executed on a computer, cause the computer to perform the method to the first aspect above.

In a third aspect of the invention, there is provided a driver guidance system comprising a taxi prediction module, and a taxi coordination module communicably coupled thereto, the taxi prediction module comprises a three-dimensional spatiotemporal model, the three-dimensional spatiotemporal model comprises a convolutional neural network trained on a first dataset of demand and supply of taxis at a plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods, wherein the three-dimensional spatiotemporal model is configured to generate a current demand count for the plurality of unit grids based on a current taxi dataset comprising current demand and supply of a taxi fleet, and the taxi coordination module is configured to provide personalised recommendations to an individual taxi driver based on the current demand count for the plurality of grid locations.

Preferably, each three-dimensional image is a $K \times K \times h$ image, wherein h represents a number of historical time period, K represents a size of a neighbourhood proximate to the unit grid, and each pixel in the three-dimensional image has dimensions $(x, y, m)$, where $(x, y)$ refer to a coordinate of the grid location and m refers to a specific time period. More preferably, data of each pixel is stored in three channels respectively encoding the number of trips originating in the coordinate of the grid location, the number of trips ending in the coordinate of

the grid location, and the number of vacant taxis observed in the coordinate of the grid location in the specific time period.

Preferably, the taxi prediction module further comprises a micro-movement model and an integrating neural network, wherein the micro-movement model is trained on a second dataset of elapsed time of most recent vacant taxi for a plurality of road links of the unit grids and summary statistics of trips in the plurality of road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region; the integrating neural network is trained to combine and assign weightage on the demand count and the micro-movement modifier to generate a combined taxi demand prediction value, wherein the taxi prediction module is further configured to generate, by the micro-movement model, a current micro-movement modifier, wherein the current micro-movement modifier reflects a current probability of taxi demand in the road links based on the current taxi dataset and to generate, by the integrating neural network, a current combined taxi demand prediction value based on the current demand count and the current micro-movement modifier; and wherein the taxi coordination module is configured to provide personalised recommendations to the individual taxi driver based on the current demand count for the plurality of grid locations includes the taxi coordination module is configured to provide personalised recommendations to the individual driver based on the current combined taxi demand prediction value for the plurality of grid locations. More preferably, the micro-movement model comprises at least two stacked Long Short-Term Memory (LSTM) networks.

Preferably, the taxi prediction module further comprises a hybrid LSTM-TCN model, the hybrid LSTM-TCN model comprises at least two LSTM networks and a temporal convolutional network (TCN) trained on the first dataset and a third dataset of exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect the demand and supply of taxis, wherein the integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier, wherein the taxi prediction module is further configured to generate, by the hybrid LSTM-

TCN model, a current exogenous taxi demand modifier based on the current taxi dataset and additional dataset comprising current exogenous factors, and to generate, by the integrating neural network, the current combined taxi demand prediction value based on the current demand count, the current micro-movement modifier and the current exogenous taxi demand modifier. More preferably, the current taxi dataset and additional dataset are fed into the at least two LSTM networks and outputs from the LSTM networks serve as inputs to the TCN to output the exogenous taxi demand modifier. In an example, the exogenous factors are selected from the following: meteorological conditions in the time period, temporal data in the time period and taxi-related data in the time period.

Preferably, the driver guidance system further comprises a display unit to receive and display the personalised recommendations from the taxi coordination module.

In a third aspect of the invention, there is provided a display unit for a driver guidance system. The display unit comprises a receiver configured to receive personalised recommendations from the driver guidance system according to the second aspect and a graphical user interface to show the personalised recommendations.

Preferably, the display unit further comprises a GPS module to determine a location of the display unit and a transmitter configured to transmit the location to the driver guidance system.

In a fourth aspect of the invention, there is provided a computer-implemented method for displaying a predicted taxi demand on a display unit. The method comprises receiving, by the display unit, a current demand count or a current combined taxi demand prediction value, wherein the current demand count or the current combined taxi demand prediction value is generated by a system implementing the method according to any of claims 1 to 8; and displaying, by the display unit, the current demand count or the current combined taxi demand prediction value.

Preferably, the method further comprises transmitting, by the display unity, a current location of the display unit to the system, wherein the system further identifies at least one unit grid or road link with taxi demand proximal to the current taxi location based on the current demand count or the current combined taxi demand prediction value; and receiving data regarding the at least one unit grid or road link.

More preferably, the method further comprises receiving and displaying, by the display unit, a personalised recommendation for the display unit, wherein the personalised recommendation is generated by the system based on the current demand count or the current combined taxi demand prediction value and location of vacant taxis.

In a fifth aspect of the invention, there is provided a method of building a taxi demand prediction model. The method comprises providing a map of a city as a plurality of unit grids; and training a convolutional neural network using a first dataset of demand and supply of taxis at the plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods.

Preferably, the method further comprises training at least two stacked Long Short-Term Memory (LSTM) networks using a second dataset of elapsed time of most recent vacant taxi for a plurality of road links of the unit grids and summary statistics of trips in the plurality of road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and to output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region; and training an integrating neural network to combine and assign weightage on the demand count and the micro-movement modifier to output a combined taxi demand prediction value.

Preferably, the method further comprises training at least two LSTM networks and a temporal convolutional network (TCN) on the first dataset and a third dataset of

exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect the demand and supply of taxis; and wherein the integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier.

**Description of Figures**

Figure (FIG.) 1 shows a map of a region, e.g. Singapore, divided into or provided by a plurality of unit grids;

Figure 2 shows a high-level design of a $M^2$-CNN model, composed of 3 major components, according to some embodiments of the invention;

Figure 3 shows: (a) a sample 2D image for grid $(x, y)$ at time period m; (b) a 3D image for grid $(x, y)$ over 4 time periods, where the temporal dimension is the vertical axis; (c) the convolutional network;

Figure 4 shows the hybrid LSTM-TCN model;

Figure 5 shows the micro-movement model;

Figure 6 shows the performance comparison of the $M^2$-CNN model against other methods from prior literature. Comparisons were performed under different demand profiles: low, medium, and high, referring to the percentile of demands at below 25%, 25% – 75%, and above 75% respectively.

Figure 7 shows the data in Figure 6 in percentages as the advantage of $M^2$-CNN model over the other prior literature methods;

Figure 8 shows the vacant roaming time performance for DMVST-NET and $M^2$-CNN;

Figure 9 shows the performance of different variants of the embodiments described herein at different demand profiles;

Figure 10 shows an example of a graphical user interface provided by a Driver Guidance System.

**Detailed Description of Embodiments of the Invention**

In the following description, numerous specific details are set forth in order to provide a thorough understanding of various illustrative embodiments of the invention. It will be understood, however, to one skilled in the art, that embodiments of the invention may be practiced without some or all of these specific details. Embodiments described in the context of one of the methods or devices are analogously valid for the other methods or devices. Similarly, embodiments

described in the context of a method are analogously valid for a device, and vice versa.

Modifications, additions, or omissions may be made to the systems, apparatuses, and methods described herein without departing from the scope of the disclosure. For example, the components of the systems and apparatuses may be integrated or separated. Moreover, the operations of the systems and apparatuses disclosed herein may be performed by more, fewer, or other components and the methods described may include more, fewer, or other steps. Additionally, steps may be performed in any suitable order. As used in this document, “each” refers to each member of a set or each member of a subset of a set.

The term “taxi” as used herein refers to a vehicle used to transport passengers and/or goods for a fare. The fare may be pre-paid or post-paid. The vehicle includes one with a driver both in vehicle and remote controlled, driverless or autonomous vehicles.

The term “current” is used to distinguish the implementation of the method from the training of the component models.

The methods and systems describe herein aim to provide good passenger demand predictions, both immediate and in the near future, to be effective. The methods and systems described herein are focussed on taxi demand prediction and use deep neural networks. The methods and systems herein aim to predict taxi demand from all channels (e.g. street hails, taxi queue, booking) compared to most existing methods which predict taxi demand for either street hail or booking. A taxi queue may be considered an organised form of a street hail and may be prevalent in certain areas where traffic restrictions are imposed on the stopping of vehicles.

A few important features of the methods and systems described herein include:

(1) The incorporation of vacant taxi’s movement traces which were not previously incorporated in predicting taxi or ride-hailing demands. This enables two important benefits:

a) Hidden information may be extracted from vacant roaming taxis, even for stretches of roads where there are currently no observed demands. This component is shown to generate the most significant impact in raising the prediction quality of the model described herein.

b) The design of the methods and systems described herein is road-segment-based, which allows accurate prediction of demand that originate from street hails or taxi queues, while also capturing pre-booked demands. This is in direct contrast to the frameworks described in references [1] and [2], and other ride-hailing service providers, which focus only on pre-booked demands.

(2) In references [1] and [2], the authors focus on capturing the impacts of demand occurrences in “related” regions on the current region at the current time; the definition of being “related” could be spatially or temporally based, but in either case, these relationships need to be specifically defined by the model builder (which may require a lot of manual-tuning and domain knowledge). In reference [1], the model builder needs to create the correlation between spatially distant regions by looking for the similarity of demand patterns. In reference [2], the model builder looks into 3 ways to correlate regions: (1) proximity, (2) similarity in the point-of-interest in the regions of interest, and (3) transportation connectivity. In contrast, a tuning-free framework that automatically includes demand/supply observations from all nearby (or proximate) regions of the selected unit grid during the most recent time periods is used in the methods and systems described herein. Advantageously, this avoids inherent bias from the model builder and requiring the model builder to have expert domain knowledge of the region (or require the input of an additional expert with the relevant domain knowledge),

The methods and systems described herein (or may be referred to as $M^2$-CNN model) is designed to predict taxi demands based on both macroscopic and microscopic information sources. Data may be taken from taxi movement logs which may be collected and stored by the operator of the taxi fleet. This data includes where and when a trip commences and ends (and in between may be assumed to be a vacant taxi). Additional data like a map of the city, weather conditions, etc, may be sourced from other available databases. The $M^2$-CNN model is a composite deep neural network that integrates multiple views utilising both real-time macroscopic and microscopic data. Compared against past approaches that utilise deep neural

networks in predicting taxi or ride-hailing demands, a tuning-free framework to directly incorporate the spatial and temporal dependencies is adopted in $M^2$-CNN model. Further, a hybrid of Long Short-Term Memory Network and Temporal Convolutional Network to incorporate real-world time series with long sequences is proposed and a unique microscopic component that attempts to extract demand-generation-related insights revealed by vacant roaming taxis is used in the $M^2$-CNN model.

The effectiveness of the $M^2$-CNN model was validated in two stages. Firstly, the approach was validated using a large-scale real-world taxi dataset containing detailed moving logs of more than 20,000 taxis and 12 million trips per month over a three-month period in Singapore. With this dataset, it was shown that the $M^2$-CNN is competitive against a wide array of approaches from prior literature and is shown in Figures 6 to 8. By analysing the effectiveness of individual components, the inclusion of microscopic information may be the most critical in generating high-quality prediction (Figure 9). The second stage of the validation is a real-world field trial, where the demand prediction engine is integrated with a data streaming service which continuously provides current locations and status of all taxis in Singapore. Based on a published demand and supply balancing algorithm [4], a Driver Guidance System (DGS) was created, which generates recommendations on where to roam for vacant taxis. With highly accurate demand prediction, it was shown that taxi drivers can reduce their vacant roaming time by 34% when they follow the recommendations.

As in most past works on taxi (or ride-hailing) demand prediction, both the spatial and temporal dimensions were discretized using fixed interval sizes. For the spatial dimension, the unit grid region was defined to be 1km by 1km; the unit grid regions are mutually exclusive, and collectively they cover all the city areas to generate demand predictions for. For the temporal dimension, the unit time period was defined to be 15-minutes long.

The unit grid $n$ was denoted as $l_n$ and let the set $L = \{l_1, \dots, l_n, \dots, l_N\}$ be the collection of all unit grids. Figure 1 shows a map of Singapore divided into or provided by a plurality of unit grids of 1km by 1km (other types of grid shapes or

size may be used as desired). Based on this grid definition, realistic geographical feature such as travelling distances and cost between grid regions may be calculated. The term “city” is not to be limited to an area of a specific population size and may refer to any urban area where a fleet of taxis may operate. Some countries and cities may have unique characteristics, for example certain areas of the cities may be restricted to specific taxi fleets, or the city may overlap with neighbouring cities or towns to form a larger urban area (e.g. a conurbation or metropolis). The grid may thus need adaption to the specific city.

Similarly, the time period $m$ was denoted as $t_m$ and let the set $T = \{t_1, \dots, t_m, \dots, t_M\}$ be the collection of all time periods. In addition, not all unit grids may be used if there are no road links within the unit grid, for example water areas like sea or reservoir, an offshore island not connected to the mainland, forested area. Almost all features included in the $M^2$-CNN model are aggregated into a particular $(n, m)$ tuple (grid $l_n$, time period $t_m$).

The framework contains three major components, as illustrated in Figure 2, and is described in greater detail next.

<u>Component (1): 3D-CNN: 3D-Spatio-Temporal Convolutional Neural Network</u>

In the deep-learning-based taxi demand prediction literature, the spatial relationships are usually captured by using a convolutional neural network (CNN) (e.g., see [1]). The basic idea is to treat the demand prediction problem as an image recognition problem, where each pixel stores demand-related information of a grid region in its red, green, and blue (RGB) channels.

The critical design decision of this approach is on what grid regions to include in the image for each grid (or unit grid) $l_n$. The most straightforward design is to include all grid regions; however, this will result in a very large image, and the prediction quality, as a result, will deteriorate (as pointed out by [1]). A suitable approach is to include only the relevant grid regions. This is where domain knowledge comes into play, and there are several different ways of identifying relevant grid regions. For example, in [1], the authors apply the proximity principle, and for each grid $l_n$, all

grid regions that are within 3 units of Chebyshev distance from the grid $l_n$ are included. Their image size is thus 7 by 7. In [2], the authors have proposed another approach, which uses a connectivity graph that reflects physical express road linkage to capture the closeness of any two regions. The composed image is thus based on the strength of linkages. In both designs (which are both competitive), the authors decide to encode only demand information that comes from the same time period $t_m$.

The design of Component (1) in the present method and system incorporates spatial proximity and temporal proximity. To achieve this, the individual pixel in the CNN is three dimensional (3D), so that both spatial and temporal dimensions can be incorporated simultaneously. This is unlike the approach in reference [1], where temporal dependency is handled separately in another component.

As shown in Figure 3 the image is defined to be of 3 dimensions, $(x, y, m)$, where $(x, y)$ refers to the grid region's (or unit grid's) spatial location, and $m$ refers to the temporal dimension, indicating the number of time periods from the current time. Figure 3a shows a sample 2D image for grid $(x, y)$ at time period $m$; Figure 3b shows the 3D image for grid $(x, y)$ over 4 time periods, where the temporal dimensional is the vertical axis; and Figure 3c shows a convolutional neural network.

For each (unit) grid $l_n$, we construct an $K \times K \times h$ image to encode all information on demand and supply counts. The parameter $h$ represents how far into the past we would want to include, and $K$ specifies the size of the included neighborhood around (or proximate to) the grid $l_n$. In one example, $h = 16$ and $K = 9$ (i.e., all regions that are within 4 units of Chebyshev distance are included).

For a pixel $(x, y, m)$ in the 3D image, the information related to the grid region $(x, y)$ in time $t_m$ is stored in channels R, G, and B. In an example, for the channel R, the number of trips originating from $(x, y)$ is encoded; for the channel G, the number of trips ending in $(x, y)$ is encoded; finally, for the channel B, the number of vacant taxis observed in $(x, y)$ during $t_m$ is encoded. This allows the model to recognise demand occurrence patterns using image recognition techniques based on the CNN

and to output a demand count based on a current taxi dataset. When the 3D-CNN model is the only component model used, the demand count would be the final taxi demand prediction value. When the 3D-CNN model is used in combination with one or both component models described herein, the 3D-CNN demand count output may be considered a demand count modifier instead as it would only be one part of the final combined taxi demand prediction value generated.

The current taxi dataset may contain all information related to the fleet of taxis in the city or a specific operator. This includes demand (trip counts), supply (number of vacant taxis), and micro-movement (elapsed time since last vacant taxi's visit at monitored road links explained further below) and may be updated every time period. As an example, the time period may be 1 minute, or any other length of time. The supply of taxis refers to taxis that are currently vacant, hence for taxis whose trips end within the grid region and become vacant would be counted as supply after the passengers alight. Thus, the current taxi dataset may include the current taxi fleet related information as described above including the current supply and demand of a taxi fleet, in other words vacant and occupied taxis in the fleet and their locations. The taxi fleet may belong to one or more operators.

This information is generated from the real-time location/status updates of all taxis and is independent of the demand prediction engine. For example, the taxis may be equipped with a Global Positioning System (GPS) transceiver to transmit its location to the fleet control centre. This information source is connected to the demand prediction engine and will be populated to all 3D images belonging to unit grids that predictions are being made for. As a result, the 3D-CNN model will automatically generate the module outputs (e.g. demand counts) for the required unit grids periodically, for example every 1 minute or any other length of time.

<u>Component (2): LSTM-TCN: A Hybrid Model of Long Short-Term Memory and Temporal Convolutional Networks</u>

The 3D-CNN component handles all information directly related to the demand occurrence and taxi supply. For other exogenous time series that could potentially have an impact on the demand occurrence (as listed in Table 1 below), these are

incorporated using a differently designed component. Table 1 provides some exogenous factors that may affect the demand and supply of taxis.

| Feature Type | Feature Description |
|---|---|
| Meteorological | • Average temperature in the time interval<br>• Average UV index in a time interval<br>• Average humidity in a time interval<br>• Cumulative rainfall in a time interval |
| Temporal | • Hour of day<br>• Minute of hour<br>• Day of week<br>• Day of month<br>• Whether the day is a weekday<br>• Whether the day is a holiday (non-working day which may vary between countries or even cities) |
| Taxi Related | • Number of trips starting in the target region<br>• Number of trips starting in the immediate neighbours of the target region<br>• Number of trips ending in the target region<br>• Number of trips ending in the immediate neighbours of the target region<br>• Number of free taxis |

Table 1: Exogeneous time series incorporated in Component (2)

To handle sequential data, the neural network model of choice is usually the Recurrent Neural Network (RNN). However, a recent study by [3] demonstrates that the Temporal Convolutional Network (TCN) sometimes outperforms RNN in certain sequence modelling tasks.

According to the methods and systems herein, a hybrid model that combines both the RNN and TCN is proposed, to take advantage of the strengths of both methods

as shown in Figure 4. As a large number of data sources is incorporated, the features to be included tend to be rather noisy. The RNN models work well with these noises and a RNN architecture, the Long Short-Term Memory (LSTM) network, is used to perform encoding and automatic feature selection. The TCN then makes use of the encoded features to model the taxi demands as a sequence. The hybrid LSTM-TCN model may be considered as a component that quantifies the impact of exogenous factors on taxi demands and is trained to generate one or more exogenous taxi demand modifiers.

All RNNs include a chain of repeating modules of neural networks. The LSTM networks also include this chain-like structure. A unit in a typical LSTM network contains four unique components: the cell, the input gate, the forget gate, and the output gate. While the state is kept in the cell, the three gates control the flow of information.

LSTM is designed to learn sequential correlations by keeping track of a cell state $c_m$ at each time interval $m$. At each time interval $m$, the LSTM requires the following inputs:

- $x_m$: the actual input data encoded as a vector.
- $h_{m-1}$: the hidden states from the previous time interval $m-1$.
- $c_{m-1}$: the cell state from the previous time interval $m-1$.

All the inputs are first aggregated to the cell through the input gate $i_m$. LSTM also has a forget gate $f_m$, and if it is activated, it can forget some previous cell $c_{m-1}^{l}$. Finally, the output gate $o_m$ controls the output of cell. The version of our LSTM implementation is as follows:

$$
\begin{aligned}
i_m &= \sigma\left(W_{ii}\, x_m + W_{hi} h_{m-1} + b_i\right), \\
f_m &= \sigma\left(W_{if}\, x_m + W_{hf} h_{m-1} + b_f\right), \\
o_m &= \sigma\left(W_{io}\, x_m + W_{ho}\, h_{m-1} + b_o\right), \\
\tilde{c}_m &= \tanh\left(W_{i\tilde{c}}\, x_m + W_{h\tilde{c}} h_{m-1} + b_{\tilde{c}}\right), \\
c_m &= f_m \circ c_{m-1} + i_m \circ \tilde{c}_m, \\
h_m &= o_m \circ \tanh(c_m).
\end{aligned}
$$

In the above implementation, $W_{ig}$, $W_{hg}$, and $b_g$ ($g \in \{i, f, o, \tilde{c}\}$) are all parameters that are to be learned.

After the input $x_m$ is sent through two LSTM layers, the sequence of the latent representations, $(h_{m-k}, \dots, h_m)$, are considered as the inputs to the TCN. The TCN uses the Dilated Convolutions to accommodate an exponentially large receptive field. The dilated convolution operation $F$ on the element $s$ of the sequence is defined as:

$$F(s) = (\boldsymbol{h} *_d f)(s) = \sum_{i=0}^{k-1} f(i)\, \boldsymbol{h}_{s-d\cdot i}$$

Where *d* is the dilation factor, *k* is the filter size, and $(s - d \cdot i)$ accounts for the direction of the past. We denote $\boldsymbol{h} \in \mathbb{R}^n$ as an one-dimensional sequence input. The operator $*_d$ is a *d*-dilated convolution, and $f: \{0, \dots, k-1\} \to \mathbb{R}$ is a filter. In our implementation, we set k to 128, and use dilation factors *d* = 1, 2, 4, 8, 16, 32, 64.

The current taxi dataset may further include current exogenous factors and along with the taxi fleet related information (see Table 1 for examples) may be used to generate the current exogenous taxi demand modifier that quantifies the impact of the current exogenous factors on the current taxi demand. In addition to the taxi fleet information, temporal data may be retrieved from a database taking into account whether it is a working or non-working day, or the computer implementing the model. Meteorological data may be similarly retrieved from a database provided by the local meteorological agency. As an example, the current exogenous factors dataset may include exogenous data from the past 16 time periods (4 hours). The length and number of time periods may be adjusted as required.

Component (3): $M^2$: Micro-Movement Model

A unique design of the methods and systems herein is the incorporation of vacant taxis' microscopic movement. This model is introduced to capture the hidden information in vacant taxis' movement: when a vacant taxi enters and exits a road link without a status change (i.e. the taxi remains vacant), it implies that no street-hail demands are observed along that road link. At the road link level, an extension to this observation is the strong positive correlation between the time elapsed since the last visit by a vacant taxi (i.e. the most recent vacant taxi) and the likelihood that the next incoming taxi would discover a demand.

To incorporate this insight into the methods and systems herein, road links that are worth monitoring were first identified (for example, only road links that generate at least 600 demands per month were monitored; in aggregate these road links generate around 70% of all street-hail demands). After identifying these road links, the arrival of vacant taxis to these links were monitored and the elapsed time since the last visit by a vacant taxi (i.e. the most recent vacant taxi) was updated (the elapsed time increases as time progresses, but resets when a vacant taxi arrives).

For the region in interest, elapsed times since the most recent vacant taxi of all monitored links in this region are collected and together with the summary statistics (for example the mean, quantiles, and variance of recent elapsed time observations for the monitored road links) of recent elapsed time observations are sent to two stacked LSTM layers as shown in Figure 5. The sequence of the latent representations $(h_{m-k}, \dots, h_m)$ were extracted as features for the fully connected layer. For a given road segment: the likelihood of a taxi seeing demand is correlated to "how long we have not seen a vacant taxi". A new feature "elapsed time since last vacant taxi" was defined as a new feature and was extracted from the data set. Other suitable neural networks may be used if desired.

The taxi related fleet information in the current taxi dataset may be used as input to the micro-movement model, and may include the data on the current elapsed time of most recent vacant taxi for the road links and current summary statistics of trips in the road links are used as the input.

The micro-movement model is thus able to correlate a probability of taxi demand in the (monitored) road links to the elapsed time since the last vacant taxi (i.e., the most recent vacant taxi) and outputs the micro-movement modifier/s on the demand counts. The micro-movement modifier outputs the contribution by the micro-movement module to the grand demand prediction model (or final combined taxi demand prediction value) and reflects the probability (i.e. likelihood) of current taxi demand in the road links.

<u>Integration of Three Component Models – Integrating Neural Network</u>

The outputs from each of the component models described above may be joined to form a tensor and fed to a 2-layer fully connected neural network, and finally to a sigmoid layer to get the final taxi demand prediction value. The taxi demand prediction value is a scaled value between 0 and 1 and needs to be scaled back to obtain the actual demand prediction. The scaling works by using the predetermined minimum and maximum of demand counts for a particular grid. The minimum and maximum demand counts are inferred from the historical dataset.

Fully connected layers are used to fuse all three models together. We fuse output $u_m^l$ from the 3D-CNN model, $v_m^l$ from the hybrid LSTM-TCN model, and $w_m^l$ from the Micro-Movement model to form a tensor:

$$z_m^l = u_m^l \oplus v_m^l \oplus w_m^l$$

We feed $z_m^l$ to the 2-layer fully connected network, and finally to a sigmoid layer to get the final prediction value $\hat{y}_{m+1}^l$:

$$\acute{y}_m^l = ReLU\left(W_{f_1}\, z_m^l + b_{f_1}\right),$$

$$\tilde{y}_m^l = ReLU\left(W_{f_2}\, \acute{y}_m^l + b_{f_2}\right),$$

$$\hat{y}_{m+1}^l = \sigma\left(W_{f_3}\, \tilde{y}_m^l + b_{f_3}\right),$$

where $W_{f_1}, W_{f_2}, W_{f_3}, b_{f_1}, b_{f_2}, b_{f_3}$ are learnable parameters, and $\sigma$ denotes the *sigmoid* activation function. The final prediction value $\hat{y}_{m+1}^l$ is in [0, 1], which will have to be scaled back to the actual counts by using the minimum and maximum of grid $l$.

In an embodiment, all three component models are used. In another embodiment, only the 3D-CNN model is used, while the integrating neural network may not be required. In another embodiment, the 3D-CNN model may be combined with either the LSTM-TCN hybrid model or the micro-movement model and the integrating neural network may accordingly be trained to combine the two respective outputs and assign the weightage accordingly.

When only the 3D-CNN component model is used, the output of the 3D-CNN would be the final taxi demand prediction value. However, when the other component models are used in combination with the 3D-CNN component model, during the training process, the output (a single predicted count) from each component model is compared to the actual count (from data in the training set), if there are any

differences (i.e. errors), the errors will be back-propagated into all layers of the integrating neural network and the connected component models to update the weights and to train the integrating neural network and the connected component models to generate the combined (final) taxi demand prediction value.

As described above, the current taxi dataset is used to generate the taxi demand prediction value (either by the 3D-CNN component model or in combination with the other component models). In addition, the current taxi dataset may be added to the existing training set to further train and update the component models and integrating neural network as appropriate.

**Empirical Evaluation**

Methods and systems described herein were evaluated against a wide range of alternatives, using detailed movement logs of around 20,000 taxis in Singapore for a three-month period. The competing approaches are listed below:

- **Historical Average (HA)**: The historical average of the demand values is employed to predict the demand value in the next time interval.
- **Linear Regression (LR)**: A classical statistical approach to model the linear relationship between a scalar response (or dependent variable) and one or more explanatory variables.
- **Support Vector Regression (SVR)**: A regression version of the Support-Vector Machine.
- **XGBoost**: A widely used gradient boosting framework.
- **Multi-Layer Perceptron (MLP)**: A class of feedforward artificial neural network (ANN).
- **Auto-Regressive Integrated Moving Average with Weekday/Weekend Indicator (ARIMA)**: A generalise ARMA model that is widely used for time series prediction tasks.
- **Deep Multi-View Spatial-Temporal Networks (DMVST-Net)**: This is a deep learning-based approach that uses multiple views of data to model and predict taxi demand [1]. CNN is used for spatial features, LSTM is used for temporal features, and Graph Embedding is used to capture semantic features. These three views are then integrated to produce the final prediction. DMVST-Net is

one of the most recent state-of-the-art deep-learning-based method in ride-hailing demand prediction.

To assess the impact of better demand prediction in this type of application, a driver guidance system (DGS) was adopted following the literature [4], and use a highly realistic microscopic taxi operation simulation [5] to evaluate the resulting gains in the social welfare (computed as all taxi's income). The design of the DGS aims to optimize the sum of: 1) the immediate movement cost, 2) the expected future revenue, and 3) the expected future movement cost. Better demand prediction improves the accuracy in estimating "the expected future revenue", thus allowing the DGS to generate recommendations with higher quality.

A set of simulations were performed to quantify the magnitude of the social welfare improvement that is achievable via better demand predictions. More specifically, to compare $M^2$-CNN against DMVST-Net [1], which is the state-of-the-art approach from the literature.

<u>Simulation Design</u>

The following common assumptions were made when setting up the simulation experiments:

1. It was assumed ALL taxis in the simulation will follow the generated guidance; according to [4], having all taxis following the guidance generates greatest social welfare gains. This assumption allows estimation of the upper bound on the gains that could result from improving demand predictions.
2. The Singapore map as shown in Figure 1 was used to set up the simulation. To provide an appropriate granularity, the grid of 1km-by-1km was defined to be the minimal geographic unit for demand and supply predictions, and the target of the recommendation. Based on this grid definition, realistic geographical features such as traveling distances and cost between grid regions were calculated.
3. The demand patterns in the simulation were derived from the historical dataset that covered weekdays. The historical dataset was divided into two parts: the training set, which contained 80% of days, and the testing set, which contained 20% of days. The demand prediction engines were trained

using only the training set, while during the actual simulation, the testing set was used to generate the actual demands. When training demand prediction engines, the time period was set to be 5-minute long, for the horizon of 6 time periods (30 minutes).

4. To account for the fact that the historical dataset did not capture unfulfilled demands, the following procedures were used to generate emulated "real" demands (which should include both fulfilled and unfulfilled demands in practice):

   a) For each grid $l$ and each time period $t$ (30 minutes), the mean ($\mu_{l,t}$) and the standard deviation ($\sigma_{l,t}$) of demand counts were first computed, based on testing datasets.

   b) The simulated demand counts ($d_{l,t}$) were sampled from the normal distribution $Normal(1.2 \cdot \mu_{l,t}, \sigma_{l,t})$ (assuming that unfulfilled demands are around 20% of the mean demands; and we apply a filter to ensure that the sampled count stays positive).

   c) The actual trips were sampled from historical trips that originate in the grid $l$ during the time $t$. Repeated sampling from the pool of historical trips for $d_{l,t}$ times was done to generate all simulated demands.

5. The guidance produced by the DGS indicates a recommended zone, e.g. unit grid, a taxi should stay in. The actual movements along the streets were decided by the historical frequency: when reaching a road intersection, the simulator sampled from the historical frequency (assumed that the choice selection followed the logistic distribution, fitted to the historical movement data) on which road segment to turn onto. The constraint during the street-level movement was that the choice should ensure that the grid-level decision is maintained (i.e., whenever possible, the guided taxis should only choose among road links that are within the recommended zone).

In the empirical comparisons, three popular error measures (RMSE, MAPE, and SMAPE) were used and are defined as follows:

- **Root-mean-square error (RMSE)**: $\sqrt{\sum_{i=1}^{N}(p_i - o_i)^2 / N}$, where $N$ is the number of data points, and $o_i$ and $g_i$ are the $i^{th}$ predicted and observed values respectively.
- **Mean absolute percentage error (MAPE)**: $\frac{1}{N}\sum_{i=1}^{N}\left|\frac{p_i - o_i}{o_i}\right|$, where $N$, $p_i$, and $o_i$ follow the same definitions as in RMSE. Unlike RMSE, which puts focus on the actual error values, MAPE quantifies errors in relative term, as the average absolute percentage errors over observed values.
- **Symmetric mean absolute percentage error (SMAPE)**: $\frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - o_i|}{(|p_i| + |o_i|)/2}$, defined similarly as MAPE, SMAPE differs a little bit in that its denominator is the average of the absolute values of both predicted and observed values. This helps to eliminate the impact of outliers, or observations with small values.

The performance comparisons using the same training and testing datasets are summarized below. Note that for all performance metrics, there were three demand scenarios: low, medium, and high, which referred to the demand percentiles at below 25%, 25% – 75%, and above 75% respectively. As shown in the evaluations, the $M^2$-CNN model outperforms all competing approaches regardless of demand scenarios.

To demonstrate the importance of incorporating the micro-movement information of vacant taxis, evaluations were conducted using variants of the approach, which included different components.

In Figure 6, it may be seen that the $M^2$-CNN model described herein has the lowest of the three error measures compared to all the other competing approaches and outperforms even the recently developed DMVST-Net approach. It may be seen that the improvements provided by $M^2$-CNN increases from low to medium to high demand and the improvements are most significant at high demand.

Figure 7 shows the performance advantage of $M^2$-CNN over each of the literature methods in percentage terms. It may be observed that the $M^2$-CNN model outperforms DMVST-NET in SMAPE by 0.1%, 2.5% and 6% in the low, medium

and high demand scenarios respectively. The performance of $M^2$-CNN against DMVST-NET is higher under the other error measures as shown in Figure 7 and against all the other tested methods.

Eight simulation instances were executed in total, and the results are summarised in Figure 8. Each randomly generated demand instance was used to test both DMVST-NET and $M^2$-CNN to ensure a fair comparison. Figure 8 shows the vacant roaming time performance for DMVST-NET and $M^2$-CNN. It may be observed that by controlling every factor and only varying the demand prediction engine, $M^2$-CNN outperforms DMVST-NET in vacant roaming time by an average of 9.4%. By examining the results in greater detail, the advantage of $M^2$-CNN over DMVST-NET is not homogeneous: the major improvement of vacant roaming occurs for instances with longer roaming times. For example, at 25% and 50% percentiles, the advantages of $M^2$-CNN over DMVST-NET were only 2.8% and 4.8%. However, for high percentiles, e.g., at 75% or at the maximum, the reductions were 8% and 19.4% respectively.

These results demonstrate that $M^2$-CNN is effective in preventing guided drivers from experiencing long roaming time. This is consistent with the much smaller standard deviation for $M^2$-CNN, which indicates that the quality of service would be higher and more stable for the guided drivers under $M^2$-CNN than DMVST-NET.

In Figure 9, “3D-CNN” refers to Component (1), “LSTM-TCN” refers to Component (2), and “Micro” refers to Component (3). As shown above, individually speaking, Component (3) is the most important in improving prediction qualities in all cases. Also worth noting is that the inclusion of Component (2) might not be that beneficial (row 2) until Component (3) is also included (row 4, the complete $M^2$-CNN).

The demand prediction methods and systems described may be incorporated into a Driver Guidance System (DGS) as described in the literature [4]. The DGS may comprise a taxi prediction module (or engine) and a taxi coordination module communicably coupled thereto. The DGS may be used to balance taxi demand and supply in real-time and may provide personalised recommendations which are driver-specific and may account for non-DGS drivers as well.

The taxi prediction module or engine may employ the $M^2$-CNN model described above. The taxi coordination module or engine is to solve a multi-period, multi-agent (each agent equals a driver) coordination problem, where the objective is to maximize the sum of all driver's revenues. The major components of the objective function are: 1) expected revenue, which is determined by the available demand and competing supply in each region, 2) immediate movement costs (if a taxi is instructed to move to another region), and 3) future expected movement cost (after a taxi reaches its recommended regions). Both "guided" and "un-guided" drivers may be considered if the latter information is available. Other similar algorithms may be used if desired to ensure optimal allocation of the taxi drivers to different predicted demand spots.

The DGS may further comprise a display unit or device communicably coupled to the taxi prediction module and/or taxi coordination module. The display unit or device may be a mobile device the driver may use and may download a mobile device application to access the DGS and display the personalised recommendations from the DGS on a graphical user interface, and/or display the taxi coordination module.

The display unit may comprise a receiver configured to received personalised recommendations from the DGS, in particular the taxi coordination module, and a graphical user interface to show the personalised recommendations. The display unit may further include a GPS transceiver to determine its location and a transmitter to transmit the location to the DGS. This allows the taxi driver to provide the current taxi location in the current taxi dataset that is used as input to the models described above. As a result, the taxi prediction module and/or the taxi coordination module may be further configured to identify at least one unit grid or road link proximal to the current taxi location based on the current demand count or the current combined taxi demand prediction value from the combination of two or more of the component models described above. This allows the taxi coordination module to generate the personalised recommendation for an individual taxi drive according to the demand and supply balancing algorithm to maximise the benefit for all or most of the taxi drivers. The application may be configured to capture or track the driver's movement

which may be used to calculate the compliance level of drivers. At the trip level, the compliance of DGS usage by the driver during all trips allows the determination of whether a trip contributed to the use of DGS. In the empirical evaluation, a trip is labelled as DGS-assisted if during the vacant roaming period right before the driver fetched the trip, the driver follows the DGS guidance for greater than 60% of the time.

Figure 10 shows an example of a mobile device showing the personalised recommendations. The DGS may show the map at different levels of enlargement (or zoom). Figure 10 on the left shows the graphical user interface displaying adjacent zone – at region level demands relative to the driver's current location when the driver is far away (outside the recommended region). Figure 10 on the right shows the graphical user interface displaying adjacent region – at street level demands relative to the driver's current location when the drive is nearby (within the recommended region).

The DGS was deployed in two field trials. In the first field trial in Singapore running for more than 1 year, more than 500 sign ups from drivers were obtained with about 50 active users. It was found that by following the personalised guidance provided by the DGS, approximately 34% less roaming time was recorded on average (11.8 mins vs 7.8 mins) and the DGS was found to effective in all hours (demands fluctuate between low, medium and high, for example depending on the time and day of the week). A second field trial was conducted in Tokyo with 29 dedicated drivers. Following DGS guidance, the drivers had approximately 12% less roaming time (17.2 mins vs 15.2 mins). Even though the second field trial was conducted when there were government advisories limiting travel and movement, the trial proved to be effective in reducing the roaming time even under challenging conditions. The shorter roaming time has financial implications as an approximately 10% increase in vacancy percentage leads to 614 yen decrease in average fare per hour.

Alternatively, the $M^2$-CNN demand prediction model and recommendation system may be integrated into a software-as-a-service product or platform and provide to taxi fleet operators to optimise their fleet usage and maximise their drivers' income.

In addition, it may be possible for the $M^2$-CNN demand prediction model and system to be deployed in autonomous vehicles (AV) fleets that allows the AV-based service fleet to be repositioned better in anticipation of future demands.

The methods, system and apparatus described herein may be employed to predict taxi demand in any city and with the driver guidance system may help alleviate imbalances in the demand and supply of taxis in the city. This is beneficial to both the consumer and driver as it minimises the wait time and roaming time respectively. Furthermore, by reducing the vacant roaming time, the unnecessary pollution and wastage of fuel caused by the vacant roaming taxis may be reduced.

In various embodiments, modules or software can be used to practice certain aspects of the invention. For example, software-as-a-service (SaaS) models or application service provider (ASP) models may be employed as software application delivery models to communicate software applications to clients or other users. Such software applications can be downloaded through an Internet connection, for example, and operated either independently (e.g., downloaded to a laptop or desktop computer system) or through a third-party service provider (e.g., accessed through a third-party web site). In addition, cloud computing techniques may be employed in connection with various embodiments of the invention. In certain embodiments, a “module” may include software, firmware, hardware, or any reasonable combination thereof.

Moreover, the processes associated with the present embodiments may be executed by programmable equipment, such as computers. Software or other sets of instructions that may be employed to cause programmable equipment to execute the processes may be stored in any storage device, such as a computer system (non-volatile) memory. Furthermore, some of the processes may be programmed when the computer system is manufactured or via a computer-readable memory storage medium.

It can also be appreciated that certain process aspects described herein may be performed using instructions stored on a computer-readable memory medium or

media that direct a computer or computer system to perform process steps. A computer-readable medium may include, for example, memory devices such as diskettes, compact discs of both read-only and read/write varieties, optical disk drives, and hard disk drives. A computer-readable medium may also include memory storage that may be physical, virtual, permanent, temporary, semi-permanent and/or semi-temporary.

A “computer,” “computer system,” “computing apparatus,” “component,” or “computer processor” may be, for example and without limitation, a processor, microcomputer, minicomputer, server, mainframe, laptop, personal data assistant (PDA), wireless e-mail device, smart phone, mobile phone, electronic tablet, cellular phone, pager, processor, fax machine, scanner, or any other programmable device or computer apparatus configured to transmit, process, and/or receive data. Computer systems and computer-based devices disclosed herein may include memory for storing certain software applications used in obtaining, processing, and communicating information. It can be appreciated that such memory may be internal or external with respect to operation of the disclosed embodiments. The memory may also include any means for storing software, including a hard disk, an optical disk, floppy disk, ROM (read only memory), RAM (random access memory), PROM (programmable ROM), EEPROM (electrically erasable PROM) and/or other computer-readable memory media. In various embodiments, a “host,” “engine,” “loader,” “filter,” “platform,” or “component” may include various computers or computer systems, or may include a reasonable combination of software, firmware, and/or hardware.

In various embodiments of the present invention, a single component may be replaced by multiple components, and multiple components may be replaced by a single component, to perform a given function or functions. Except where such substitution would not be operative to practice embodiments of the present invention, such substitution is within the scope of the present invention. Any of the servers described herein, for example, may be replaced by a “server farm” or other grouping of networked servers (e.g., a group of server blades) that are located and configured for cooperative functions. It can be appreciated that a server farm may serve to distribute workload between/among individual components of the farm and may

expedite computing processes by harnessing the collective and cooperative power of multiple servers. Such server farms may employ load-balancing software that accomplishes tasks such as, for example, tracking demand for processing power from different machines, prioritizing and scheduling tasks based on network demand, and/or providing backup contingency in the event of component failure or reduction in operability.

In general, it will be apparent to one of ordinary skill in the art that various embodiments described herein, or components or parts thereof, may be implemented in many different embodiments of software, firmware, and/or hardware, or modules thereof. The software code or specialized control hardware used to implement some of the present embodiments is not limiting of the present invention. For example, the embodiments described hereinabove may be implemented in computer software using any suitable computer programming language such as .NET, SQL, MySQL, or HTML using, for example, conventional or object-oriented techniques. Programming languages for computer software and other computer-implemented instructions may be translated into machine language by a compiler or an assembler before execution and/or may be translated directly at run time by an interpreter. Examples of assembly languages include ARM, MIPS, and x86; examples of high level languages include Ada, BASIC, C, C++, C#, COBOL, Fortran, Java, Lisp, Pascal, Object Pascal; and examples of scripting languages include Bourne script, JavaScript, Python, Ruby, PHP, and Perl. Various embodiments may be employed in a Lotus Notes environment, for example. Such software may be stored on any type of suitable computer-readable medium or media such as, for example, a magnetic or optical storage medium. Thus, the operation and behaviour of the embodiments are described without specific reference to the actual software code or specialized hardware components. The absence of such specific references is feasible because it is clearly understood that artisans of ordinary skill would be able to design software and control hardware to implement the embodiments of the present invention based on the description herein with only a reasonable effort and without undue experimentation.

References:

[1] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., Li, Z.: Deep multi-view spatial-temporal network for taxi demand prediction. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

[2] Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., Liu, Y.: Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In: Thirty-Third AAAI Conference on Artificial Intelligence (2019)

[3] Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)

[4] Shashi Shekhar Jha, Shih-Fen Cheng, Meghna Lowalekar, Nicholas Wong, Rishikeshan Rajendram, Trong Khiem Tran, Pradeep Varakantham, Nghia Truong Trong, and Firmansyah Abd Rahman. 2018. Upping theGame of Taxi Driving in the Age of Uber. In Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18), 7779–7785

[5] Shih-Fen Cheng and Thi Duong Nguyen. 2011. TaxiSim: A multiagent simulation platform for evaluating taxi fleet operations. In 2011 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 14–21.

**CLAIMS**

1. A computer-implemented method for predicting taxi demand, the method comprising

generating, by a three-dimensional spatiotemporal model, a current demand count for a plurality of unit grids based on a current taxi dataset comprising current demand and supply of a taxi fleet, wherein the three-dimensional spatiotemporal model comprises a convolutional neural network trained on a first dataset of demand and supply of taxis at the plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods;

generating, by a micro-movement model, a current micro-movement modifier, wherein the current micro-movement modifier reflects a current probability of taxi demand in a plurality of road links based on the current taxi dataset comprising current elapsed time of most recent vacant taxi for the road links and current summary statistics of trips in the road links, wherein the micro-movement model is trained on a second dataset of elapsed time of most recent vacant taxi for the road links and summary statistics of trips in the road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and to output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region; and

generating, by an integrating neural network, a current combined taxi demand prediction value based on the current demand count and the current micro-movement modifier, wherein the integrating neural network is trained to combine and assign weightage on the demand count and the micro-movement modifier.

2. The method according to claim 1 wherein each three-dimensional image is a $K \times K \times h$ image, wherein $h$ represents a number of historical time period, $K$ represents a size of a neighbourhood proximate to the unit grid, and each pixel in the three-dimensional image has dimensions $(x, y, m)$, where $(x, y)$ refer to a coordinate of the unit grid and $m$ refers to a specific time period.

3. The method according to claim 2 wherein data of each pixel is stored in three channels respectively encoding the number of trips originating in the coordinate of the unit grid, the number of trips ending in the coordinate of the unit grid, and the number of vacant taxis observed in the unit grid in the specific time period.

4. The method according to any one of claims 1 to 3 wherein the micro-movement model comprises at least two stacked Long Short-Term Memory (LSTM) networks.

5. The method according to any of claims 1 to 4 further comprising
generating, by a hybrid LSTM-TCN model, a current exogenous taxi demand modifier based on both the current taxi dataset and additional dataset comprising current exogenous factors, wherein the hybrid LSTM-TCN model comprises at least two LSTM networks and a temporal convolutional network (TCN) trained on the first dataset and a third dataset of exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect demand and supply of taxis,
wherein generating, by the integrating neural network, the current combined taxi demand prediction value includes generating, by the integrating neural network, the current combined taxi demand prediction value based on the current demand count, the current micro-movement modifier and the current exogenous taxi demand modifier,
wherein the integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier.

6. The method according to claim 5 wherein the current taxi dataset and additional dataset is provided to the at least two LSTM networks and outputs from the LSTM networks serve as inputs to the TCN to output the current exogenous taxi demand modifier.

7. The method according to any of claims 5 to 6 wherein the exogenous factors are selected from the following: meteorological conditions in the time period, temporal data in the time period and taxi-related data in the time period.

8. The method according to any of claims 1 to 7 wherein the current taxi dataset comprises a current taxi location, and the method further comprises identifying at least one unit grid or road link with taxi demand proximal to the current taxi location based on the current demand count or the current combined taxi demand prediction value.

9. The method according to claim 8 further comprising generating a personalised recommendation for an individual taxi driver based on the current demand count or the current combined taxi demand prediction value and location of vacant taxis.

10. A non-transitory computer readable medium comprising instructions which, when executed on a computer, cause the computer to perform the method according to any of claims 1 to 9.

11. A driver guidance system comprising a taxi prediction module, and a taxi coordination module communicably coupled thereto,
the taxi prediction module comprises a three-dimensional spatiotemporal model, a micro-movement model and an integrating neural network, the three-dimensional spatiotemporal model comprises a convolutional neural network trained on a first dataset of demand and supply of taxis at a plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods, wherein the three-dimensional spatiotemporal model is configured to generate a current demand count for the plurality of unit grids based on a current taxi dataset comprising current demand and supply of a taxi fleet, and the taxi coordination module is configured to provide personalised recommendations to an individual taxi driver based on the current demand count for the plurality of grid locations,

wherein the micro-movement model is trained on a second dataset of elapsed time of most recent vacant taxi for a plurality of road links of the unit grids and summary statistics of trips in the plurality of road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region;

the integrating neural network is trained to combine and assign weightage on the demand count and the micro-movement modifier to generate a combined taxi demand prediction value,

wherein the taxi prediction module is further configured to generate, by the micro-movement model, a current micro-movement modifier, wherein the current micro-movement modifier reflects a current probability of taxi demand in the road links based on the current taxi dataset and to generate, by the integrating neural network, a current combined taxi demand prediction value based on the current demand count and the current micro-movement modifier; and

wherein the taxi coordination module is configured to provide personalised recommendations to the individual taxi driver based on the current demand count for the plurality of grid locations includes the taxi coordination module is configured to provide personalised recommendations to the individual driver based on the current combined taxi demand prediction value for the plurality of grid locations.

12. The driver guidance system according to claim 11 wherein each three-dimensional image is a $K \times K \times h$ image, wherein $h$ represents a number of historical time period, $K$ represents a size of a neighbourhood proximate to the unit grid, and each pixel in the three-dimensional image has dimensions $(x, y, m)$, where $(x, y)$ refer to a coordinate of the grid location and $m$ refers to a specific time period.

13. The driver guidance system according to claim 12 wherein data of each pixel is stored in three channels respectively encoding the number of trips originating in the coordinate of the grid location, the number of trips ending in

the coordinate of the grid location, and the number of vacant taxis observed in the coordinate of the grid location in the specific time period.

14. The driver guidance system according to any one of claims 11 to 13, wherein the micro-movement model comprises at least two stacked Long Short-Term Memory (LSTM) networks.

15. The driver guidance system according to any of claims 11 to 14 wherein the taxi prediction module further comprises a hybrid LSTM-TCN model, the hybrid LSTM-TCN model comprises at least two LSTM networks and a temporal convolutional network (TCN) trained on the first dataset and a third dataset of exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect the demand and supply of taxis,
wherein the integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier,
wherein the taxi prediction module is further configured to generate, by the hybrid LSTM-TCN model, a current exogenous taxi demand modifier based on the current taxi dataset and additional dataset comprising current exogenous factors, and to generate, by the integrating neural network, the current combined taxi demand prediction value based on the current demand count, the current micro-movement modifier and the current exogenous taxi demand modifier.

16. The driver guidance system according to claim 15 wherein the current taxi dataset and additional dataset are fed into the at least two LSTM networks and outputs from the LSTM networks serve as inputs to the TCN to output the exogenous taxi demand modifier.

17. The driver guidance system according to any of claims 15 to 16 wherein the exogenous factors are selected from the following: meteorological conditions in the time period, temporal data in the time period and taxi-related data in the time period.

18. The driver guidance system according to any of claims 11 to 17 further comprising a display unit to receive and display the personalised recommendations from the taxi coordination module.

19. A display unit for a driver guidance system, the display unit comprising a receiver configured to receive personalised recommendations from the driver guidance system according to any of claims 11 to 17 and a graphical user interface to show the personalised recommendations.

20. The display unit according to claim 19 further comprising a GPS module to determine a location of the display unit and a transmitter configured to transmit the location to the driver guidance system.

21. A computer-implemented method for displaying a predicted taxi demand on a display unit, the method comprising receiving, by the display unit, a current demand count or a current combined taxi demand prediction value, wherein the current demand count or the current combined taxi demand prediction value is generated by a system implementing the method according to any one of claims 1 to 7; and displaying, by the display unit, the current demand count or the current combined taxi demand prediction value.

22. The method according to claim 21 further comprising transmitting, by the display unit, a current location of the display unit to the system, wherein the system further identifies at least one unit grid or road link with taxi demand proximal to the current location of the display unit based on the current demand count or the current combined taxi demand prediction value; and receiving data regarding the at least one unit grid or road link.

23. The method according to claim 22, receiving and displaying, by the display unit, a personalised recommendation for the display unit, wherein the personalised recommendation is generated by the system based on the current demand count or the current combined taxi demand prediction value and location of vacant taxis.

24. A method of building a taxi demand prediction model, the method comprises providing a map of a city as a plurality of unit grids;

training a convolutional neural network using a first dataset of demand and supply of taxis at the plurality of unit grids over a plurality of time periods to output a demand count, wherein the first dataset is encoded as a plurality of three-dimensional images, and each three-dimensional image represents one unit grid over the plurality of time periods;

training at least two stacked Long Short-Term Memory (LSTM) networks using a second dataset of elapsed time of most recent vacant taxi for a plurality of road links of the unit grids and summary statistics of trips in the plurality of road links to correlate a probability of taxi demand in the road links to an elapsed time of most recent vacant taxi in the road links and to output a micro-movement modifier reflecting the probability of taxi demand in the road links, wherein the road links are adjacent to form a region; and

training an integrating neural network to combine and assign weightage on the demand count and the micro-movement modifier to output a combined taxi demand prediction value.

25. The method according to claim 24 further comprises training at least two LSTM networks and a temporal convolutional network (TCN) on the first dataset and a third dataset of exogenous factors to output an exogenous taxi demand modifier, wherein the exogenous taxi demand modifier quantifies how exogenous factors affect the demand and supply of taxis; and wherein he integrating neural network is further trained to combine and assign weightage on the exogenous taxi demand modifier.