



(51) International Patent Classification:

G06N 3/08 (2006.01) G06Q 20/38 (2012.01)
G06Q 20/16 (2012.01)

(21) International Application Number:

PCT/US2021/057774

(22) International Filing Date:

02 November 2021 (02.11.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

17/087,895 03 November 2020 (03.11.2020) US

(71) Applicant: CAPITAL ONE SERVICES, LLC [US/US];
1680 Capital One Drive, McLean, Virginia 22102 (US).

(72) Inventors: CHENG, Lin Ni Lisa; c/o Capital One Services, LLC, 1680 Capital One Drive, McLean, Virginia 22102 (US). FIELDS, John; c/o Capital One Services, LLC, 1680 Capital One Drive, McLean, Virginia 22102 (US).

(74) Agent: KELLY, Scott M.; 1100 13th Street, N.W., Suite 1200, Washington, District of Columbia 20005 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,

CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: LEVERAGING BLOCKCHAIN BASED MACHINE LEARNING MODELING FOR EXPENSE CATEGORIZATION

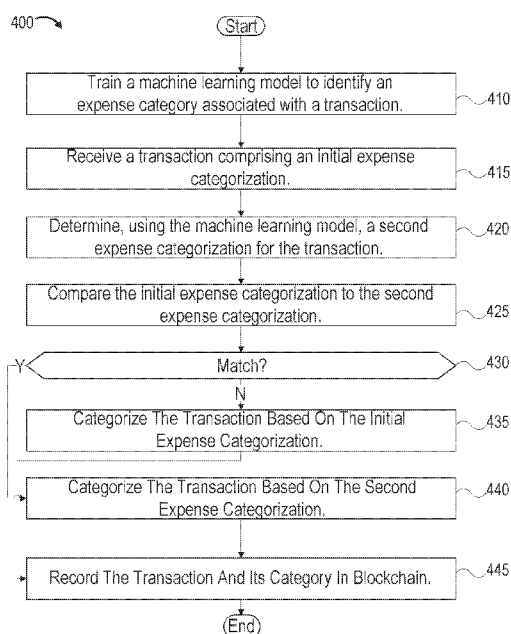


FIG. 4

(57) Abstract: Systems and methods disclosed herein provide automatic expense categorization of transactions or expenditures based on a machine learning (ML) model trained using anonymized transactional data for expenditures that are stored in a public blockchain. The anonymized transactional data for the expenditure and the expense category, may be distributed throughout the blockchain network and recorded in the blockchain. In some implementations, an expenditure may be submitted to the blockchain network for expense categorization. The transactional data for the expenditure may be anonymized to remove confidential and personal identifying information (PII) before it is distributed throughout the blockchain network. Each participating node of the blockchain network may utilize the ML model to identify an expense category for the expenditure. The participating nodes may provide a consensus mechanism in order to arrive at a shared understanding of how to categorize the expenditure.

LEVERAGING BLOCKCHAIN BASED MACHINE LEARNING MODELING FOR EXPENSE CATEGORIZATION

RELATED APPLICATIONS

- [1] This application claims priority to U.S. Application Serial No. 17/087,895, filed on November 3, 2020, whose contents are expressly incorporated herein by reference in their entirety.

FIELD OF USE

- [2] Aspects described herein generally relate to expense categorization communication, and hardware and software related thereto. More specifically, one or more aspects describe herein improve expense categorization of transactions by leveraging a machine learning model trained using anonymized transactional data stored in a public blockchain.

BACKGROUND

- [3] Business expenditures may be categorized using, for example, an accountant or accounting software. An expenditure may be categorized differently based on various criteria. As an example, a law firm paying for another law firm's services may be categorized differently than an import/exporter paying for a law firm's services. As another example, a marketing company's expenditures related to marketing services are likely to be normal operating costs rather than advertising expenses.

SUMMARY

- [4] The following presents a simplified summary of various aspects described herein. This summary is not an extensive overview, and is not intended to identify key or critical elements or to delineate the scope of the claims. The following summary merely presents some concepts in a simplified form as an introductory prelude to the more detailed description provided below. Corresponding apparatus, systems, and computer-readable media are also within the scope of the disclosure.
- [5] According to one aspect, the disclosure relates to a computer-implemented method that includes training, based on transactional data stored in a blockchain, a machine learning model to identify an expense category associated with a transaction. The method includes receiving a first transaction comprising a first expense category associated with the first transaction. The method also includes determining a second expense category for the first transaction. The second expense category is determined based on at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a

product identification associated with the first transaction, or a service identification associated with the first transaction, and by: selecting the second expense category for the first transaction based on a consensus of a plurality of participating nodes of the blockchain. The method includes determining whether the first expense matches the second expense category, categorizing, based on a determination that the first expense category matches the second expense category, the first transaction. The method includes, based on the categorizing of the first transaction, storing transactional data for the first transaction in the blockchain. The transactional data for the first transaction comprises an indication of an expense category based on the determination that the first expense category matches the second expense category.

- [6] According to another aspect, the disclosure relates to an apparatus that includes one or more processors, and a memory storing computer-readable instructions. The computer-readable instructions, when executed by the one or more processors, cause the apparatus to train, based on transactional data stored in a blockchain, a machine learning model to identify an expense category associated with a transaction, receive a first transaction comprising a first categorization associated with the first transaction, determine, using the machine learning model, a second categorization for the first transaction. The machine learning model is trained to determine the second categorization is based on at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a product identification associated with the first transaction, or a service identification associated with the first transaction. The computer-readable instructions, when executed by the one or more processors, cause the apparatus to also determine whether the first categorization matches the second categorization, categorize, based on a determination that the first categorization matches the second categorization, the first transaction; and based on the categorizing of the first transaction, store transactional data for the first transaction in the blockchain. The transactional data for the first transaction comprises an indication of an expense category based on the determination that the first categorization matches the second categorization.
- [7] According to another aspect, the disclosure relates to non-transitory machine-readable medium storing instructions that, when executed by one or more processors, cause the one or more processors to perform steps comprising training, based on transactional data stored in a blockchain, a machine learning model to identify a category associated with a transaction, and receiving a first transaction comprising a first categorization associated with the first transaction, determining, using the machine learning model, a second

categorization for the first transaction. The second categorization is based on at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a product identification associated with the first transaction and comprising a stock-identification code (SKU), a service identification associated with the first transaction, a geographic location, or a tax code. The steps include determining whether the first categorization matches the second categorization, categorizing, based on a determination that the first categorization matches the second categorization, the first transaction; and based on the categorizing of the first transaction: anonymizing the first transaction by removing personally identifying information; and based on the categorizing of the first transaction, storing transactional data for the first transaction in the blockchain. The transactional data for the first transaction comprises an indication of an expense category based on the determination that the first expense category matches the second expense category.

BRIEF DESCRIPTION OF THE DRAWINGS

- [8] The present disclosure is described by way of example and not limited in the accompanying figures in which like reference numerals indicate similar elements and in which:
- [9] FIG. 1 shows an example of a control processing system in which one or more aspects described herein may be implemented;
- [10] FIG. 2 shows an example computing device in accordance with one or more aspects described herein;
- [11] FIG. 3 shows an example system for leveraging blockchain based machine learning modeling for expense categorization; and
- [12] FIG. 4 shows a flow diagram of an example method for leveraging blockchain based machine learning modeling for expense categorization;

DETAILED DESCRIPTION

- [13] In the following description of the various embodiments, reference is made to the accompanying drawings, which form a part hereof, and in which is shown by way of illustration various embodiments in which aspects of the disclosure may be practiced. It is to be understood that other embodiments may be utilized and structural and functional modifications may be made without departing from the scope of the present disclosure. Aspects of the disclosure are capable of other embodiments and of being practiced or being carried out in various ways. In addition, it is to be understood that the phraseology and

terminology used herein are for the purpose of description and should not be regarded as limiting. Rather, the phrases and terms used herein are to be given their broadest interpretation and meaning.

- [14] Existing techniques for categorizing business expenses utilize transitional databases or data centers along with rule-based processing. These techniques do not facilitate semi-open and anonymous data sharing between businesses and service providers. Moreover, rules-based processing systems may not be accurate since they lack the flexibility to adapt quickly to changing trends based on updates to relevant parameters. Further, the rules are defined by human involvement, making these techniques predictable and easy to penetrate. Additionally, existing techniques rely on categorization provided by a financial institution or accounting systems or software which may not be able to accurately identify the expense category for an expenditure. For example, accounting software utilized by small businesses often require users to have an account in order to provide a correct expense category. While financial institutions and accounting software may accurately categorize some types of expenses, it may be challenging to categorize transactions at larger stores and chains. For example, a \$100 transaction at a restaurant may reliably be categorized as a food expense. However, a \$100 transaction at a large department store may be for clothing, or a service, such as a haircut. Additionally, categorizing the expenditures for small businesses may also be challenging since there are so many of them and it may not always be easy to determine what type of business it is.
- [15] Systems and methods disclosed herein provide automatic expense categorization of transactions or expenditures based on a machine learning (ML) model trained using anonymized transactional data for expenditures that are stored in a public blockchain. A machine learning (ML) model may be trained to evaluate an expenditure and identify an expense category for the expenditure. The training data for the ML model may be received from one or more sources. For example, one of the sources of training data may be transactional information stored or recorded in a public blockchain. Other sources may include accounting systems, accountants, financial institution databases, etc. Utilizing the ML model, an expenditure may be categorized as a legal expense, client development expense, marketing expense, food expense, travel expense, entertainment expense, etc.
- [16] As discussed above, one of the sources of training data may be transactional information stored or recorded in a public blockchain. Entities, such as businesses or individuals, may choose to record their expenditures in a public blockchain by submitting information about

their expenditures to a blockchain network. A submitted expense or expenditure may be anonymized by removing confidential and personally identifiable information (PII) before it is distributed throughout the blockchain network. The anonymized data is subsequently recorded in the blockchain. As an example, Company A may purchase a product from a merchant. Company A may submit transactional data about the purchase to the blockchain network. The submitted data may include an initial expense category for the purchase. The initial expense category may have been determined by Company A's accounting software. The transactional data associated with the purchase may include an identification of the merchant, the SKU, and/or the price. The submitted data may be anonymized by removing information, such as an identification of Company A, and other potentially confidential information. The anonymized data may be distributed to the participating nodes of the blockchain network and subsequently recorded in a blockchain stored by the blockchain network. The transactional data recorded in the blockchain may serve as the training data for the ML model.

- [17] A business may submit an expenditure for expense categorization. An expense category for the expenditure may be identified using the ML model. Expert scoring and feedback may be added to increase the level of confidence of the categorization. After an expense category is identified, the transactional data for the expenditure may be anonymized to remove confidential and personal identifying information (PII). The anonymized transactional data for the expenditure and the expense category, may be distributed throughout the blockchain network and recorded in the blockchain.
- [18] In some implementations, an expenditure may be submitted to the blockchain network for expense categorization. The transactional data for the expenditure may be anonymized to remove confidential and personal identifying information (PII) before it is distributed throughout the blockchain network. Each participating node of the blockchain network may utilize the ML model to identify an expense category for the expenditure. The participating nodes may provide a consensus mechanism in order to arrive at a shared understanding of how to categorize the expenditure.
- [19] Systems and methods disclosed herein may provide business profile specific expense categorization through machine learning based on an entity's (e.g. an individual, a merchant, an organization, a company or a business) expenditures. The confirmed transaction data and/or related smart contracts may be distributed or recorded into the blockchain network for additional security and privacy protection. Smart contracts may be developed on top of

the blockchain to execute an action for some desired outcome, either on behalf of one, more than one or all companies. For example, a categorization of an expense may be of interest to the vendor/retailer. Although this information is anonymous, in aggregate this information could be informational. With multiple inputs via, for example, a financial institution, Enterprise Resource Planning (ERP), an accounting system, regulations (i.e. legal or tax regulations), and/or user adjustments, the process may be streamlined from multiple sources for multiple verifications in order to improve the accuracy of recording the expenditure for reporting, taxation, and other legal purposes. The decentralized network distribution may enable private node networks among the users, verification and trust-proof onboarded partners, from service providers, such as, cross-border financial institutions' networks, ERP, accountants, etc., for anonymous sharing via a double-blind confirmation process and training of an ML model. The training of the ML model may enhance the alignment of the automatic process based on the business profile and past transaction activity. The blockchain node distribution may provide privacy protections and additional security measurements in place to prevent manipulation of stored records or data. Anonymizing data prior to storing it in the blockchain enables users and regulators to drill down into the transactional data beyond the aggregated level and gain meaningful insight into the data. Additional ML models may be implemented for converting stored transactional data into synthetic data prior to extraction of the data from the blockchain.

- [20] Systems and methods disclosed herein may improve the speed and accuracy with which computers recognize and categorize business expenses, especially eclectic expenses, such as local shops and non-chain retailers (e.g., boutique retailers, local restaurants, etc.). Systems and methods disclosed herein may enable businesses to categorize expenditures with a better accuracy than that of a human, such as an accountant. Thus, implementations improve the speed and accuracy with which computers categorize business expenses. The ML based expense categorizing may also include mapping to different tax or legal codes, thus, enabling businesses to streamline their tax processes. Expense categorization based on considering various tax codes may be helpful for businesses operating across different countries.
- [21] Additionally, implementations enable a business to safely and securely automate accounting and legal processes via crowd sourced knowledge from other businesses without exposing its own identity, trade secrets, and/or other confidential information. The confidence level of the process may be measured and improved via a feedback loop. Implementations also

provide standardization that improves the consistency of expense categorization among businesses having the same profile (industry).

- [22] FIG. 1 shows a system 100. The system 100 may include at least one client device 110, at least one database system 120, and/or at least one server system 130 in communication via a network 140. It will be appreciated that the network connections shown are illustrative and any means of establishing a communications link between the computers may be used. The existence of any of various network protocols such as TCP/IP, Ethernet, FTP, HTTP and the like, and of various wireless communication technologies such as GSM, CDMA, WiFi, and LTE, is presumed, and the various computing devices described herein may be configured to communicate using any of these network protocols or technologies. Any of the devices and systems described herein may be implemented, in whole or in part, using one or more computing systems described with respect to FIG. 2.
- [23] Client device 110 may access server applications and/or resources using one or more client applications (not shown) as described herein. Client device 110 may be a mobile device, such as a laptop, smart phone, or tablet, or computing devices, such as a desktop computer or a server. Alternatively, client device 110 may include other types of devices, such as game consoles, camera/video recorders, video players (e.g., incorporating DVD, Blu-ray, Red Laser, Optical, and/or streaming technologies), smart TVs, and other network-connected appliances, as applicable.
- [24] Database system 120 may be configured to maintain, store, retrieve, and update information for server system 130. Further, database system may provide server system 130 with information periodically or upon request. In this regard, database system 120 may be a distributed database capable of storing, maintaining, and updating large volumes of data across clusters of nodes. Database system 120 may provide a variety of databases including, but not limited to, relational databases, hierarchical databases, distributed databases, in-memory databases, flat file databases, XML databases, NoSQL databases, graph databases, and/or a combination thereof.
- [25] Server system 130 may be configured with a server application (not shown) that is capable of interfacing with client application and database system 120 as described herein. In this regard, server system 130 may be a stand-alone server, a corporate server, or a server located in a server farm or cloud-computer environment. According to some examples, server system 130 may be a virtual server hosted on hardware capable of supporting a plurality of virtual servers.

- [26] Network 140 may include any type of network. For example, network 140 may include a local area network (LAN), a wide area network (WAN), a wireless telecommunications network, and/or any other communication network or combination thereof. It will be appreciated that the network connections shown are illustrative and any means of establishing a communications link between the computers may be used. The existence of any of various network protocols such as TCP/IP, Ethernet, FTP, HTTP and the like, and of various wireless communication technologies such as GSM, CDMA, WiFi, and LTE, is presumed, and the various computing devices described herein may be configured to communicate using any of these network protocols or technologies.
- [27] The data transferred to and from various computing devices in a system 100 may include secure and sensitive data, such as confidential documents, customer personally identifiable information, and account data. Therefore, it may be desirable to protect transmissions of such data using secure network protocols and encryption, and/or to protect the integrity of the data when stored on the various computing devices. For example, a file-based integration scheme or a service-based integration scheme may be utilized for transmitting data between the various computing devices. Data may be transmitted using various network communication protocols. Secure data transmission protocols and/or encryption may be used in file transfers to protect the integrity of the data, for example, File Transfer Protocol (FTP), Secure File Transfer Protocol (SFTP), and/or Pretty Good Privacy (PGP) encryption. In many embodiments, one or more web services may be implemented within the various computing devices. Web services may be accessed by authorized external devices and users to support input, extraction, and manipulation of data between the various computing devices in the system 100. Web services built to support a personalized display system may be cross-domain and/or cross-platform, and may be built for enterprise use. Data may be transmitted using the Secure Sockets Layer (SSL) or Transport Layer Security (TLS) protocol to provide secure connections between the computing devices. Web services may be implemented using the WS-Security standard, providing for secure SOAP messages using XML encryption. Specialized hardware may be used to provide secure web services. For example, secure network appliances may include built-in features such as hardware-accelerated SSL and HTTPS, WS-Security, and/or firewalls. Such specialized hardware may be installed and configured in the system 100 in front of one or more computing devices such that any external devices may communicate directly with the specialized hardware.
- [28] Turning now to FIG. 2, a computing device 200 that may be used with one or more of the computational systems is described. The computing device 200 may include a processor

203 for controlling overall operation of the computing device 200 and its associated components, including RAM 205, ROM 207, input/output device 209, communication interface 211, and/or memory 215. A data bus may interconnect processor(s) 203, RAM 205, ROM 207, memory 215, I/O device 209, and/or communication interface 211. In some embodiments, computing device 200 may represent, be incorporated in, and/or include various devices such as a desktop computer, a computer server, a mobile device, such as a laptop computer, a tablet computer, a smart phone, any other types of mobile computing devices, and the like, and/or any other type of data processing device. Additionally or alternatively, computing device 200 may represent, be incorporated in, and/or include various devices such as a computing device associated with a node in a blockchain network.

- [29] Input/output (I/O) device 209 may include a microphone, keypad, touch screen, and/or stylus through which a user of the computing device 200 may provide input, and may also include one or more of a speaker for providing audio output and a video display device for providing textual, audiovisual, and/or graphical output. Software may be stored within memory 215 to provide instructions to processor 203 allowing computing device 200 to perform various actions. For example, memory 215 may store software used by the computing device 200, such as an operating system 217, application programs 219, and/or an associated internal database 221. The various hardware memory units in memory 215 may include volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, data structures, program modules, or other data. Memory 215 may include one or more physical persistent memory devices and/or one or more non-persistent memory devices. Memory 215 may include, but is not limited to, random access memory (RAM) 205, read only memory (ROM) 207, electronically erasable programmable read only memory (EEPROM), flash memory or other memory technology, optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium that may be used to store the desired information and that may be accessed by processor 203.
- [30] Communication interface 211 may include one or more transceivers, digital signal processors, and/or additional circuitry and software for communicating via any network, wired or wireless, using any protocol as described herein.
- [31] Processor 203 may include a single central processing unit (CPU), which may be a single-core or multi-core processor, or may include multiple CPUs. Processor(s) 203 and

associated components may allow the computing device 200 to execute a series of computer-readable instructions to perform some or all of the processes described herein. Although not shown in FIG. 2, various elements within memory 215 or other components in computing device 200, may include one or more caches, for example, CPU caches used by the processor 203, page caches used by the operating system 217, disk caches of a hard drive, and/or database caches used to cache content from database 221. For embodiments including a CPU cache, the CPU cache may be used by one or more processors 203 to reduce memory latency and access time. A processor 203 may retrieve data from or write data to the CPU cache rather than reading/writing to memory 215, which may improve the speed of these operations. In some examples, a database cache may be created in which certain data from a database 221 is cached in a separate smaller database in a memory separate from the database, such as in RAM 205 or on a separate computing device. For instance, in a multi-tiered application, a database cache on an application server may reduce data retrieval and data manipulation time by not needing to communicate over a network with a back-end database server. These types of caches and others may be included in various embodiments, and may provide potential advantages in certain implementations of devices, systems, and methods described herein, such as faster response times and less dependence on network conditions when transmitting and receiving data.

- [32] Although various components of computing device 200 are described separately, functionality of the various components may be combined and/or performed by a single component and/or multiple computing devices in communication without departing from the invention.
- [33] FIG. 3 shows an example system 300 for leveraging blockchain based machine learning modeling for expense categorization. The system 300 comprises a first server 305, a second server 310, and a database, such as the ML model 335, interconnected via a network 315. The network 315 may include a public distributed peer-to-peer (P2P) network or blockchain network 320. A memory 325 of the server 305 may include one or more applications, such as, an expense categorization application 330.
- [34] The expense categorization application 330 may train the machine learning (ML) model 335 to identify an expense category for a given transaction. The ML model 335 may be trained based on evaluating transactional data received from multiple sources. For example, the ML model 335 may be trained based on transactional data recorded and/or stored in the blockchain of the blockchain network 320. The transactional data may include information

about expenditures stored and/or recorded in the blockchain of the blockchain network 320. The expenditures may be submitted by various sources (e.g. financial institutions, individuals, companies, businesses, accountants, accounting systems, etc.). Legal and tax code information may also be provided as inputs to the ML model 335. The ML model 335 may utilize the legal and tax code information to provide different categorizations for the same expenditure due to different tax laws, regulations, and/or geographic locations. For example, the same expenditure may be categorized differently under different countries' or states' tax codes.

- [35] The ML model 335 may be trained to identify an expense category for a transaction. The ML model 335 may be trained using data and/or information stored in blockchain 320. For example, training the ML model 335 may comprise determining expense categories for transactions recorded in blockchain 320, for example, by analyzing transactional data. The transactional data may comprise information relevant for determining an expense category for the transaction, such as, a merchant associated with the transaction, a price or amount associated with the transaction, an identification of an item or product, for example, an SKU code, or an identification of a service associated with the transaction. As more and more transactions are recorded in the blockchain and subsequently included in the training set, the accuracy of the ML model's expense categorization increases.
- [36] As discussed above, an entity may submit an expenditure to be recorded in the blockchain. For example, the entity may send information about its expenditure from the second server 310 to the first server 305. The expense categorization application 330 may anonymize the transactional data associated with the expenditure to remove personal identifying information (PII), trade secrets, and/or other confidential information. The expense categorization application 330 may then submit the anonymized transactional data to the blockchain network 320. The distributed anonymized transactional data is recorded in the blockchain. The anonymized transactional data recorded in the blockchain may be utilized as a training set for the ML model 335.
- [37] A company may submit, via the second server 310, information about an expenditure to the first server 305 in order to receive an expense categorization for the expenditure. The information about the expenditure may include an initial expense categorization. The initial expense categorization may be determined by, for example, the company's accounting software. The expense categorization application 330 may anonymize certain information related to the expenditure. The expense categorization application 330 may determine, using

the ML model 335, a second expense categorization for the transaction. As discussed above, the ML model 335 may be trained to determine the second expense categorization based on information associated with the expenditure, such as, a merchant, an amount, a product identification, and/or a service identification. The product identification may include a stock-identification code (SKU), or UPC. Other factors or criteria related to the expenditure, such as a type of the business submitting the expenditure, may also be considered leading to a categorization with a higher level of confidence. For example, an expenditure may be categorized differently depending on whether the expenditure was associated with a plumbing company or a defense contractor. Profiling may be utilized and includes matching an entity's transactions with other entities' transactions. For example, the ML model 335 may indicate that the expenditure submitted by company A has been identified by other companies with the same profile as company A to be Category X.

- [38] The expense categorization may also include mapping to different tax or legal codes. An expenditure may need to be categorized differently in order to receive tax credit according to a country's tax laws. For example, under one country's tax law, only 50% of an expenditure related to entertainment may be claimed for tax credit, while in another other country, 100% of that expenditure may be claimed for tax credit. Expense categorization based on tax code considerations may be helpful for businesses operating across multiple countries.
- [39] The expense categorization application 330 may determine whether the initial categorization matches the second expense categorization. If the initial categorization matches the second categorization, then the expense categorization application 330 may output the initial categorization. If the initial categorization does not match the second expense category, the expense categorization application 330 may notify (e.g., alert) an administrator. Additionally or alternatively, the expense categorization application 330 may classify the expenditure (e.g., transaction) using the second expense categorization and output. In some implementations, the expense categorization application 330 may, based on determining the expense categorizing of the expenditure, record the relevant transactional data for the expenditure and the expense category in the blockchain. The recorded transaction data and expense category for the expenditure may be provided back to the ML model 335 via a feedback loop.
- [40] Participating nodes in the blockchain network 320 may provide a consensus mechanism in order to arrive at a shared understanding of how to categorize an expense. Referring back

to the example above, a company may submit, via the second server 310, information about an expenditure to the first server 305 in order to receive an expense categorization for the expenditure. The information about the expenditure may include an initial expense categorization. The expense categorization application 330 may anonymize information related to the expenditure, and then submit the anonymized expenditure to the blockchain network 320. Each participating node of the blockchain network 320 may categorize the submitted expenditure based on utilizing the ML model 335. For example, a first participating node of the blockchain network 320 may categorize the expenditure and come up with an expense category. If a second participating node in the blockchain network 320 categorizes the expenditure and comes up with the same expense category as that of the first participating node, then the categorization of the expenditure is confirmed by the two participating nodes. If a majority of the participating nodes come up with the same expense category, then consensus is reached. If a majority of participating nodes do not come up with the same expense category, then consensus is not reached, and additional inputs may be required. These inputs may be taken from a professional accountant or accounting software and may result in a higher score and thus, a recommended expense categorization. The recommended expense categorization may remain as the default reference point in the ML model. The user may still be able to record the information in the blockchain. However, the recorded information may include a notation of the recommendation. Additionally, participating nodes in the blockchain network 320 would not know who provided the expense categorization or who requested the expense categorization. Once the expenditure is categorized based on a consensus mechanism, the expense categorization application 330 may receive confirmed expense categorization data for the expenditure. After categorization is confirmed, the transactional data associated with the expenditure may be recorded in the blockchain.

- [41] In some cases, if there is insufficient training data or the transactional data does not include suitable information for the ML model 335, the expense categorization may vary and it may not be possible to provide a categorization with a high level of confidence. In other words, the transactions related to a specific type of expenditure may not be readily reconcilable to a specific expense category. In such cases, another participating node in the blockchain network may attempt to provide more insight into the transaction and suggest a category based on other criteria, for example, the type of the business. This participating node may be the manufacturer or retailer that would be interested in augmenting the system and may provide information to help fill in the gaps, even though this node itself is not categorizing

the expense. In this scenario, the blockchain network 320 would see only one participating node providing this information and that consensus has not yet been reached.

- [42] FIG. 4 is a flow diagram of an example method 400 for leveraging blockchain based machine learning modeling for expense categorization performed by a computing device (e.g. the expense categorization application 330 executing on the server 305, as shown in FIG. 3). Alternatively or additionally, some or all of the steps of the method 400 may be performed by one or more other computing devices. The steps of the method 400 may be modified, omitted, and/or performed in other orders, and/or other steps added.
- [43] At step 410, a computing device (e.g. the expense categorization application 330 executing on the server 305) may train a machine learning (ML) model, such as ML model 335, to identify an expense category for a given transaction. The ML model may be trained based on evaluating transactional data received from multiple sources. For example, the ML model may be trained based on transactional data recorded and/or stored in a blockchain, such as blockchain network 320. The expenditures may be submitted by various sources (e.g. financial institutions, individuals, companies, businesses, accountant, accounting systems, etc.). Legal and tax code information may also be provided as inputs to the ML model. The ML model may utilize the legal and tax code information to provide different categorization of the same expenditure due to different tax laws, regulations, or geographic locations. For example, the same expenditure may be categorized differently under different countries' tax codes.
- [44] As discussed above, an entity may submit an expenditure to be recorded in the blockchain. For example, the entity may send information about its expenditure from the second server 310 to the first server 305. The expense categorization application 330 may anonymize the transactional data associated with the expenditure to remove personal identifying information (PII), trade secrets, and/or other confidential information. The expense categorization application 330 may then submit the anonymized transactional data to the blockchain network 320. The distributed anonymized transactional data is recorded in the blockchain. The anonymized transactional data recorded in the blockchain may be utilized as a training set for the ML model 335.
- [45] The ML model 335 may be trained to identify an expense category for a transaction. Training the ML model 335 may include determining expense categories for transactions recorded in the blockchain of the blockchain network 320. The ML model 335 may be trained using supervised learning, unsupervised learning, back propagation, transfer

learning, stochastic gradient descent, learning rate decay, dropout, max pooling, batch normalization, long short-term memory, skip-gram, or any equivalent deep learning technique. The transactional data associated with a transaction recorded in the blockchain may include information relevant for determining an expense category for the transaction, such as, a merchant associated with the transaction, a price or amount associated with the transaction, an identification of an item or product, for example, an SKU code or UPC, or an identification of a service associated with the transaction. As more and more transactions are recorded in the blockchain and subsequently included in the training set, the accuracy of the ML model's expense categorization increases.

- [46] At step 415, a computing device (e.g. the expense categorization application 330 executing on the server 305) may receive a request for categorization of a transaction. For example, a company may submit, via the second server 310, information about an expenditure to the first server 305 in order to receive an expense categorization for the expenditure. The information about the expenditure may include an initial expense categorization. The initial expense categorization may be determined by, for example, the company's accounting software. The expense categorization application 330 may anonymize certain information related to the expenditure.
- [47] At step 420, a computing device (e.g. the expense categorization application 330 executing on the server 305) may determine, using the ML model 335, a second categorization for the transaction. The second categorization may be based on at least one of a merchant associated with the transaction, an amount associated with the transaction, a product identification associated with the transaction, and/or a service identification associated with the transaction. The product identification may include a stock-identification code (SKU) or UPC. Other factors or criteria related to the expenditure, such as a type of the business, may also be considered leading to a categorization with a higher level of confidence. For example, an expenditure may be categorized differently depending on whether the expenditure was associated with a plumbing company or a defense contractor. Profiling may be utilized and includes matching an entity's transactions with other entities' transactions. For example, the ML model 335 may indicate that the expenditure submitted by company A has been identified by other companies with the same profile as company A to be Category X.
- [48] At step 425, a computing device (e.g. the expense categorization application 330 executing on the server 305) may compare the initial categorization to the second categorization. At

step 430, a computing device (e.g. the expense categorization application 330 executing on the server 305) may determine that the initial categorization does not match the second categorization and the process 400 proceeds to step 435. At step 435, a computing device (e.g. the expense categorization application 330 executing on the server 305) may categorize the transaction, based the second categorization. If the initial categorization does not match the second expense category, the expense categorization application 330 may notify (e.g., alert) an administrator. However, at step 430, if a computing device (e.g. the expense categorization application 330 executing on the server 305) determines that the initial categorization matches the second categorization, then the process 400 proceeds to step 440. At step 440, a computing device (e.g. the expense categorization application 330 executing on the server 305) may categorize the transaction, based on the initial categorization.

- [49] At step 445, a computing device (e.g. the expense categorization application 330 executing on the server 305) may, based on the categorizing of the transaction, store or record the transactional data for the transaction in the blockchain of the blockchain network 320. The transactional data for the transaction may include an indication of the expense category based on the determination that the initial expense category matches the second expense category. The determined expense category may also be sent from the first server 305 to the second server 310 and to the entity that submitted the request for the expense categorization.
- [50] The method and systems described herein provide and enable businesses to categorize expenditures with an accuracy close to what it would be using an accountant and at a significantly lower cost than using an accountant. Thus, implementations improve the speed and accuracy with which computers categorize business expenses. Additionally, mapping the business profile would provide better quality of results. With multiple inputs via, for example, a financial institution, Enterprise Resource Planning (ERP), an accounting system, regulations (i.e. legal or tax regulations), and user adjustments, the process may be streamlined from multiple sources for multiple verifications in order to improve the accuracy of recording the expenditure for reporting, taxation, and other legal purposes. The decentralized network distribution may enable private node networks among the users, verification and trust-proof onboarded partners, from service providers, such as, cross-border financial institutions' networks, ERP, accountants, etc., for anonymous sharing via a double-blind confirmation process and training of an ML model. The training of the ML model may enhance the alignment of the automatic process based on the business profile and past transaction activity. The blockchain node distribution may provide privacy protections and additional security measurements in place to prevent manipulation of stored

records or data. Anonymizing data prior to storing it in the blockchain enables users and regulators to drill down into the transactional data beyond the aggregated level and gain meaningful insight. Additional ML models may be implemented for converting stored transactional data into synthetic data prior to extraction of the data from the blockchain.

[51] One or more aspects discussed herein may be embodied in computer-usable or readable data and/or computer-executable instructions, such as in one or more program modules, executed by one or more computers or other devices as described herein. Generally, program modules include routines, programs, objects, components, data structures, and the like that perform particular tasks or implement particular abstract data types when executed by a processor in a computer or other device. The modules may be written in a source code programming language that is subsequently compiled for execution, or may be written in a scripting language such as (but not limited to) HTML or XML. The computer executable instructions may be stored on a computer readable medium such as a hard disk, optical disk, removable storage media, solid-state memory, RAM, and the like. As will be appreciated by one of skill in the art, the functionality of the program modules may be combined or distributed as desired in various embodiments. In addition, the functionality may be embodied in whole or in part in firmware or hardware equivalents such as integrated circuits, field programmable gate arrays (FPGA), and the like. Particular data structures may be used to more effectively implement one or more aspects discussed herein, and such data structures are contemplated within the scope of computer executable instructions and computer-usable data described herein. Various aspects discussed herein may be embodied as a method, a computing device, a system, and/or a computer program product.

[52] Although the present invention has been described in certain specific aspects, many additional modifications and variations would be apparent to those skilled in the art. In particular, any of the various processes described above may be performed in alternative sequences and/or in parallel (on different computing devices) in order to achieve similar results in a manner that is more appropriate to the requirements of a specific application. It is therefore to be understood that the present invention may be practiced otherwise than specifically described without departing from the scope and spirit of the present invention. Thus, embodiments of the present invention should be considered in all respects as illustrative and not restrictive. Accordingly, the scope of the invention should be determined not by the embodiments illustrated, but by the appended claims and their equivalents.

WHAT IS CLAIMED IS:

1. A computer-implemented method comprising:
 - training, based on transactional data stored in a blockchain, a machine learning model to identify an expense category associated with a transaction;
 - receiving a first transaction comprising a first expense category associated with the first transaction;
 - determining a second expense category for the first transaction, wherein the second expense category is selected based on a consensus of a plurality of participating nodes of the blockchain and is determined using at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a product identification associated with the first transaction, or a service identification associated with the first transaction;
 - determining whether the first expense matches the second expense category;
 - categorizing, based on a determination that the first expense category matches the second expense category, the first transaction; and
 - based on the categorizing of the first transaction, storing transactional data for the first transaction in the blockchain, wherein the transactional data for the first transaction comprises an indication of an expense category based on the determination that the first expense category matches the second expense category.
2. The method of claim 1, further comprising:
 - anonymizing, prior to the storing the transactional data for the first transaction in the blockchain, the first transaction by removing personally identifiable information (PII).
3. The method of claim 1, wherein the product identification comprises a stock-identification code (SKU).
4. The method of claim 1, wherein the training the machine learning model comprises:
 - determining, by one or more nodes of the plurality of participating nodes of the blockchain, a category for a second transaction, wherein the determination is based on at least one of: a second merchant associated with the second transaction, a second amount associated with the second transaction, a product identification associated with the second transaction, or a service identification associated with the second transaction; and

storing second transactional data for the second transaction in the blockchain, wherein the second transactional data comprises the category for the second transaction determined by the one or more nodes of the plurality of participating nodes of the blockchain.

5. The method of claim 4, wherein determining the category for the second transaction comprises:

selecting the category for the second transaction based on a consensus of the plurality of participating nodes of the blockchain.

6. The method of claim 1, wherein the consensus is based on a consensus model that comprises at least one of: a proof of work consensus model, a proof of stake consensus model, or a proof of authority consensus model.

8. The method of claim 1, wherein the determining the second categorization for the first transaction is based on a geographic location.

9. The method of claim 1, wherein the determining the second categorization for the first transaction is based on a tax code.

10. The method of claim 1, further comprising:

receiving a second transaction comprising a third categorization associated with the second transaction;

determining, using the machine learning model, a fourth categorization for the second transaction, wherein the fourth categorization is based on at least one of: a merchant associated with the second transaction, an amount associated with the second transaction, a product identification associated with the second transaction, or a service identification associated with the second transaction;

determining whether the third categorization matches the fourth categorization; and

notifying, based on a determination that the third categorization does not match the fourth categorization, an administrator.

11. The method of claim 10, further comprising:
 - creating, based on a determination that the third categorization does not match the fourth categorization, a new category for the second transaction, wherein the new category is based on the fourth categorization; and
 - categorizing, using the new category the second transaction.
12. An apparatus comprising:
 - one or more processors; and
 - memory storing instructions that, when executed by the one or more processors, cause the apparatus to:
 - train, based on transactional data stored in a blockchain, a machine learning model to identify an expense category associated with a transaction;
 - receive a first transaction comprising a first categorization associated with the first transaction;
 - determine, using the machine learning model, a second categorization for the first transaction, wherein the machine learning model is trained to determine the second categorization based on a consensus of a plurality of participating nodes of the blockchain and based on using at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a product identification associated with the first transaction, or a service identification associated with the first transaction;
 - determine whether the first categorization matches the second categorization;
 - categorize, based on a determination that the first categorization matches the second categorization, the first transaction; and
 - based on the categorizing of the first transaction, store transactional data for the first transaction in the blockchain, wherein the transactional data for the first transaction comprises an indication of an expense category based on the determination that the first categorization matches the second categorization.
13. The apparatus of claim 12, wherein the instructions, when executed by the one or more processors, cause the apparatus to:
 - anonymize, prior to the storing the first transaction in the blockchain, the first transaction by removing personally identifiable information (PII).

14. The apparatus of claim 12, wherein the product identification comprises a stock-identification code (SKU).
15. The apparatus of claim 12, wherein the training the machine learning model comprises:
determining, by one or more nodes of a plurality of participating nodes of the blockchain, a category for a second transaction, wherein the determination is based on at least one of: a second merchant associated with the second transaction, a second amount associated with the second transaction, a product identification associated with the second transaction, or a service identification associated with the second transaction; and
storing second transactional data for the second transaction in the blockchain, wherein the second transactional data comprises the category for the second transaction determined by the one or more nodes of the plurality of participating nodes of the blockchain.
16. The apparatus of claim 15, wherein the determining the category for the third transaction comprises:
selecting the category for the third transaction based on a consensus of the plurality of participating nodes of the blockchain.
17. The apparatus of claim 16, wherein the consensus is based on a consensus model that comprises at least one of: a proof of work consensus model, a proof of stake consensus model, or a proof of authority consensus model.
18. The apparatus of claim 12, wherein the determining the second categorization for the first transaction is based on a geographic location.
19. The apparatus of claim 12, wherein the determining the second categorization for the first transaction is based on a tax code.
20. A non-transitory machine-readable medium storing instructions that, when executed by one or more processors, cause the one or more processors to perform steps comprising:
training, based on transactional data stored in a blockchain, a machine learning model to identify a category associated with a transaction;
receiving a first transaction comprising a first expense category associated with the first transaction;

determining, using the machine learning model, a second expense category for the first transaction, wherein the second expense category is selected based on a consensus of a plurality of participating nodes of the blockchain and is determined using at least one of: a merchant associated with the first transaction, an amount associated with the first transaction, a product identification associated with the first transaction and comprising a stock-identification code (SKU), a service identification associated with the first transaction, a geographic location, or a tax code;

determining whether the first expense matches the second expense category;

categorizing, based on a determination that the first expense category matches the second expense category, the first transaction; and

based on the categorizing of the first transaction:

anonymizing the first transaction by removing personally identifying information; and

based on the categorizing of the first transaction, storing transactional data for the first transaction in the blockchain, wherein the transactional data for the first transaction comprises an indication of an expense category based on the determination that the first expense category matches the second expense category.

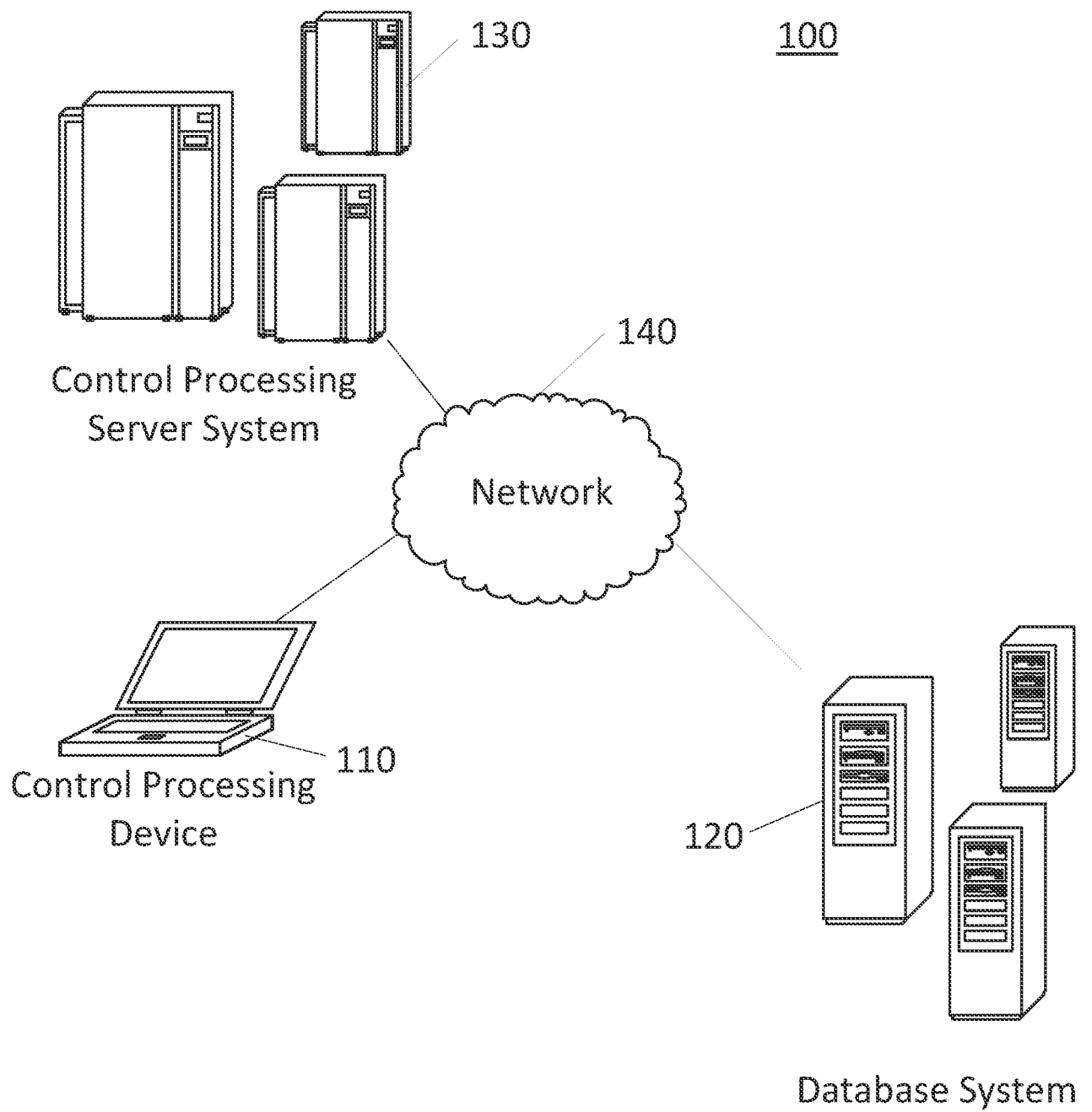


FIG. 1

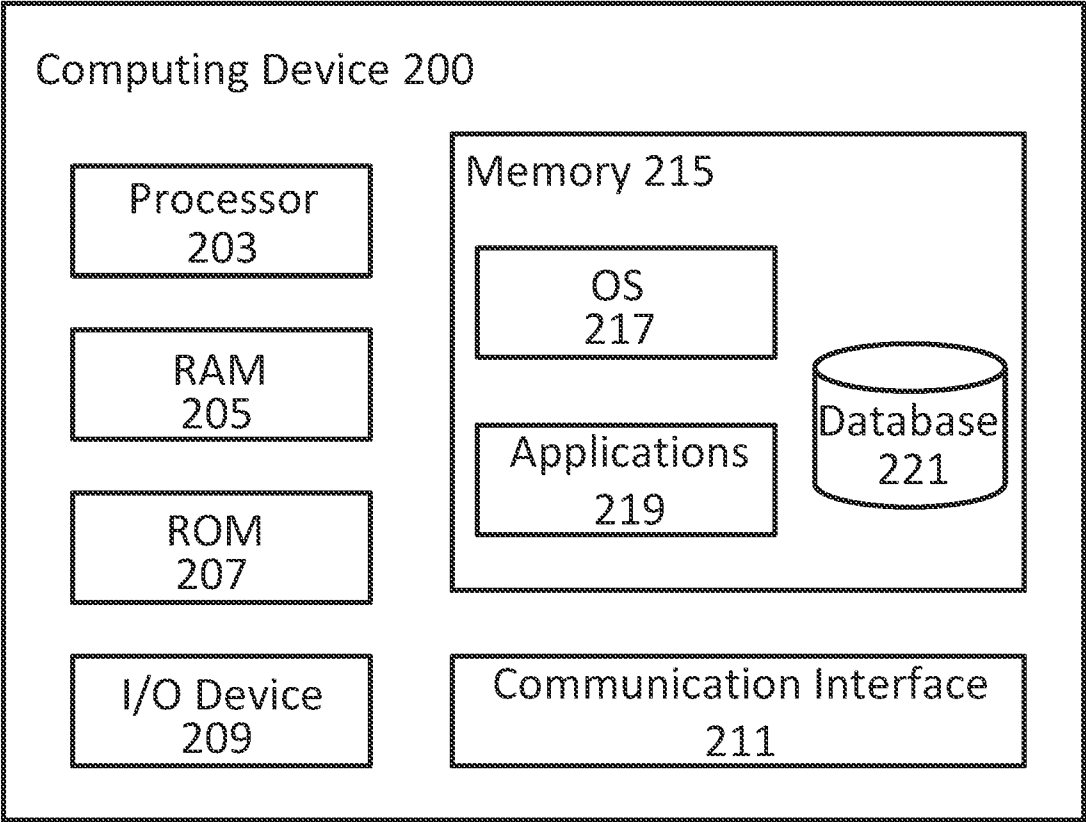


FIG. 2

300

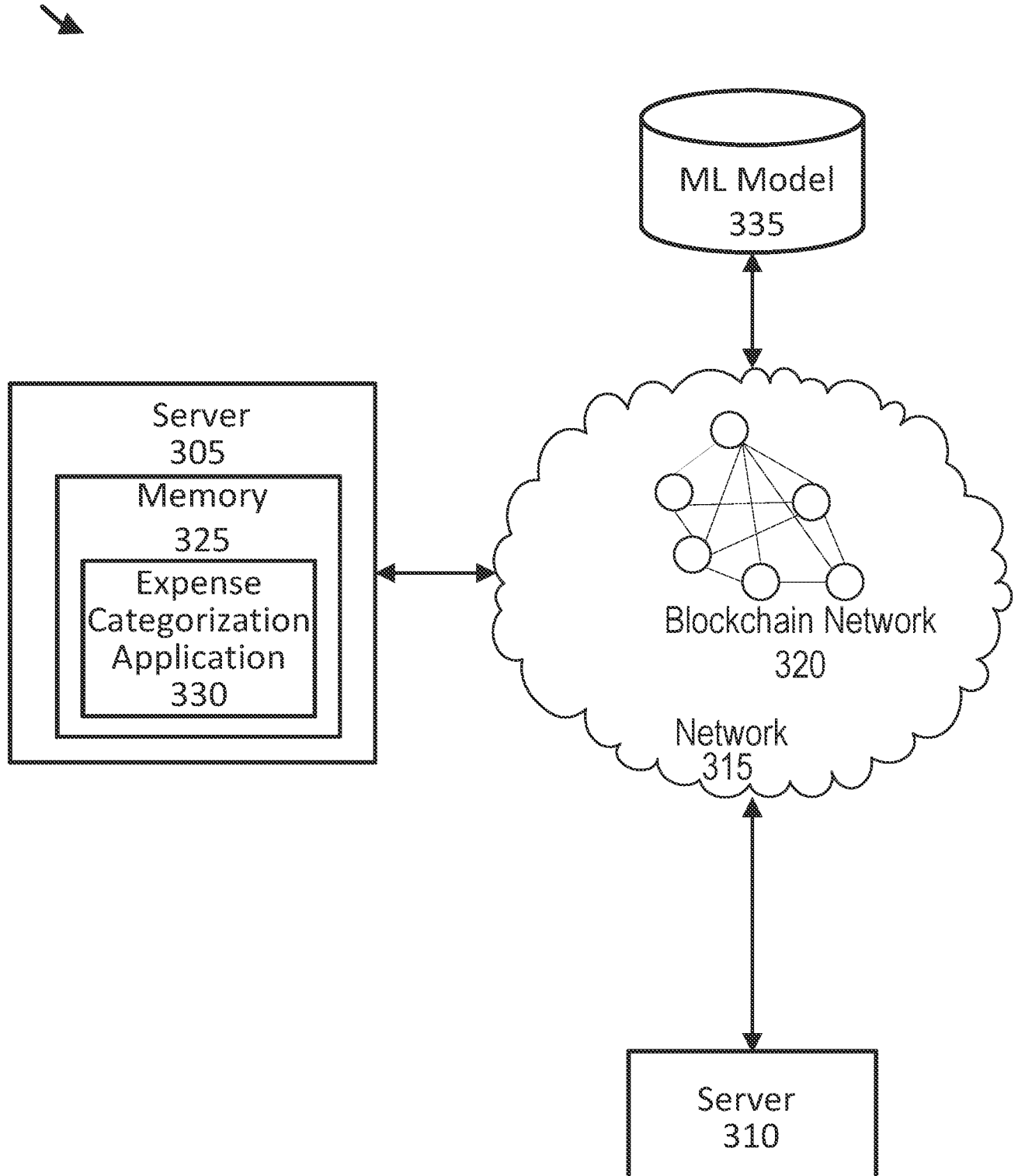


FIG. 3

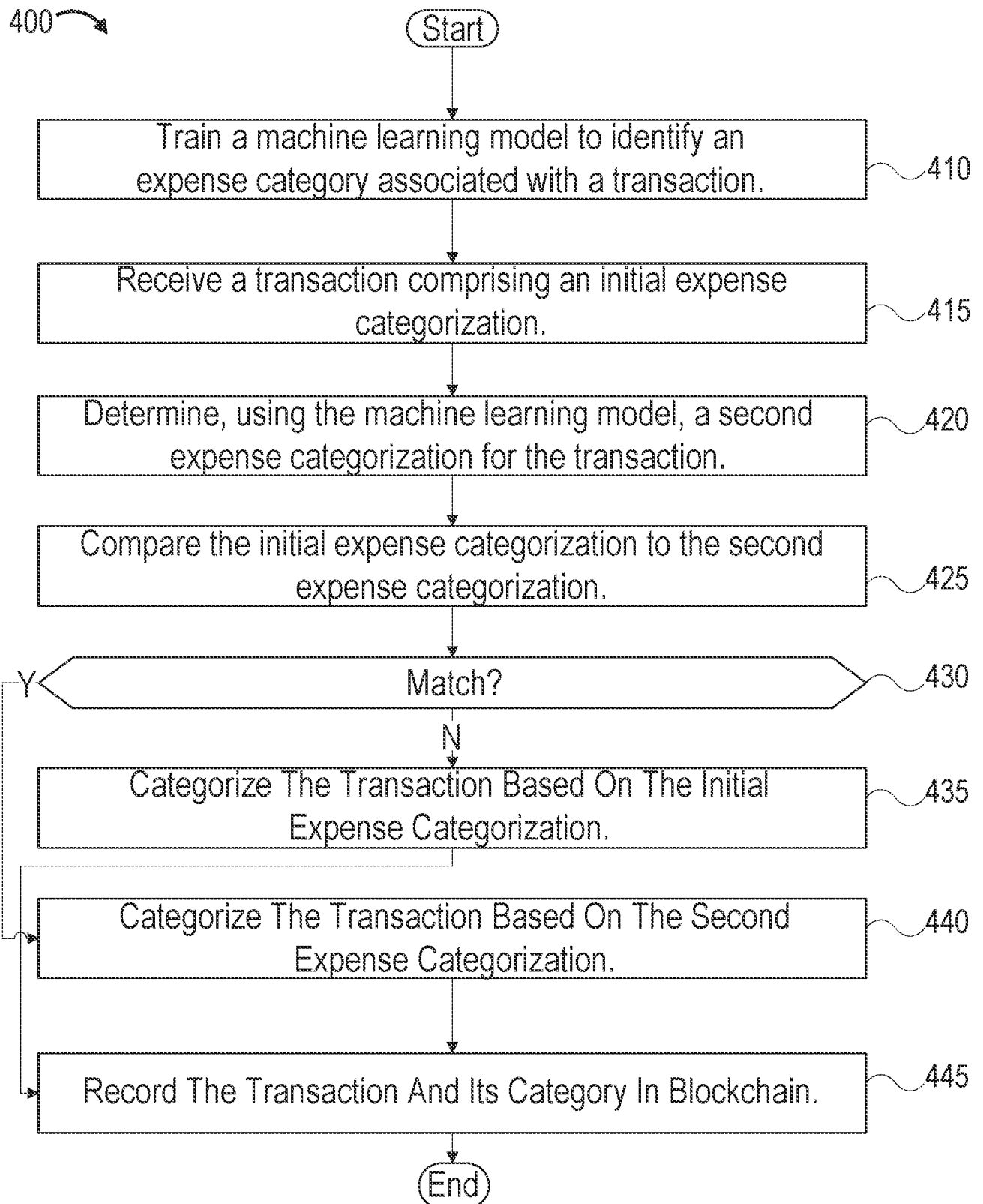


FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US21/57774

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G06N 3/08; G06Q 20/16; G06Q 20/38 (2021.01)

CPC - G06N 3/08; G06Q 20/16; G06Q 20/389; H04L 9/0618; H04L 9/0637; G06Q 20/14; G06Q 20/223; G06Q 20/382; G06Q 20/401

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016/0078561 A1 (SPRINGHEAD, INC.) 17 March 2016; entire document	1-20
A	WO 2018/064645 A1 (3DPP, LLC) 05 April 2018; entire document	1-20
A	US 2015/0227890 A1 (BEDNAREK, K ET AL.) 13 August 2015; entire document	1-20
A	US 2005/0278221 A1 (HAHN-CARLSON, D ET AL.) 15 December 2005; entire document	1-20

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

03 January 2022 (03.01.2022)

Date of mailing of the international search report

JAN 21 2022

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

Telephone No. PCT Helpdesk: 571-272-4300