(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 January 2025 (23.01.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/017100 A1

(51) International Patent Classification:
G06V 10/764 (2022.01) *G06V 10/82* (2022.01)
G06N 3/042 (2023.01) *G16H 30/40* (2018.01)
G06T 7/00 (2017.01) *G06V 10/774* (2022.01)
G06V 10/776 (2022.01)

(21) International Application Number:
PCT/EP2024/070334

(22) International Filing Date:
17 July 2024 (17.07.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/527,578 18 July 2023 (18.07.2023) US

(71) Applicant: **DEEPMIND TECHNOLOGIES LIMITED** [GB/GB]; 5 New Street Square, London EC4A 3TW (GB).

(72) Inventors: **DOUCET, Arnaud**; 6 Pancras Square, London N1C 4AG (GB). **CEMGIL, Ali Taylan**; 6 Pancras Square, London N1C 4AG (GB). **STUTZ, David**; 6 Pancras Square, London N1C 4AG (GB).

(74) Agent: **FISH & RICHARDSON P.C.**; HighLight Business Towers, Mies-van-der-Rohe-Str. 8, 80807 München (DE).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE,

(54) Title: CONFORMAL PREDICTION USING AMBIGUOUS CALIBRATION EXAMPLES

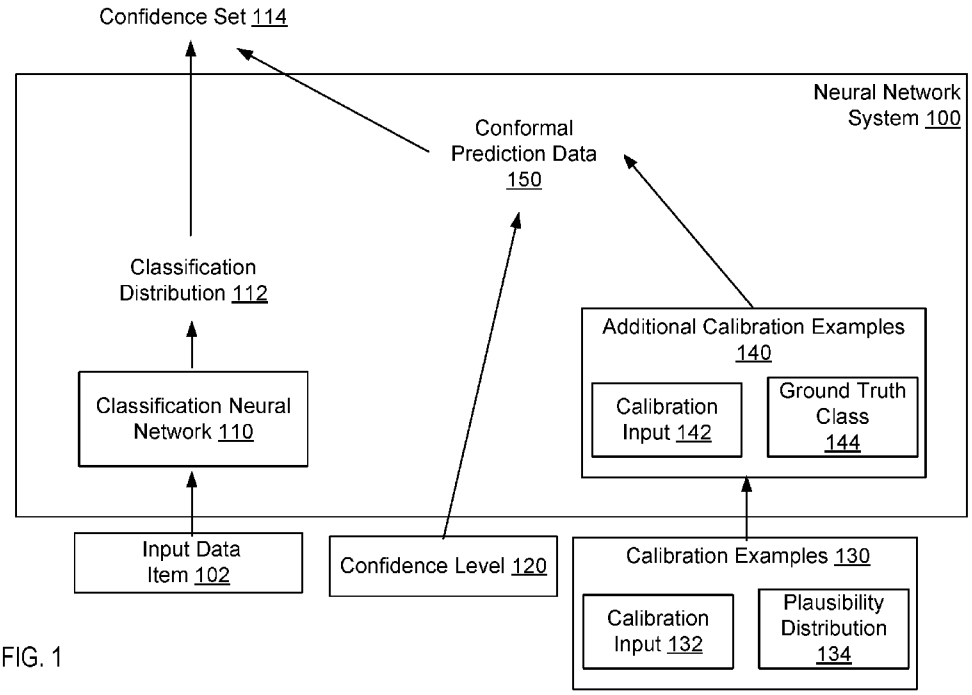


FIG. 1

(57) Abstract: Methods, systems, and apparatus, including computer programs encoded on computer storage media, for conformal prediction using ambiguous calibration examples. In particular, how to perform conformal prediction using calibration examples that include plausibility distributions is described.

[Continued on next page]

WO 2025/017100 A1

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))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

CONFORMAL PREDICTION USING AMBIGUOUS CALIBRATION EXAMPLES

CROSS-REFERENCE TO RELATED APPLICATION

This application claims priority to U.S. Provisional Application No. 63/527,578, filed on July 18, 2024. The disclosure of the prior application is considered part of and is incorporated by reference in the disclosure of this application.

BACKGROUND

This specification relates to processing inputs using machine learning models to perform classification.

As one example, neural networks are machine learning models that employ one or more layers of nonlinear units to predict an output for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to another layer in the network, e.g., the next hidden layer or the output layer. Each layer of the network generates an output from a received input in accordance with current values of a respective set of weights.

SUMMARY

This specification describes a system implemented as computer programs on one or more computers that performs conformal prediction using ambiguous calibration examples.

In one aspect, a method includes obtaining a plurality of calibration examples for calibrating an image classification neural network for conformal prediction, wherein each calibration example comprises (i) a calibration image and (ii) a plausibility distribution for the calibration image that comprises a respective score for each of a plurality of classes, and wherein the image classification neural network has been trained to process an input that comprises an input image to generate as output a classification distribution that comprises a respective classification score for each of the plurality of classes; obtaining a confidence level that indicates a required confidence in conformal predictions generated by the image classification neural network; for each calibration example, generating a plurality of additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising, for each additional calibration example: obtaining the respective additional calibration image from the calibration image in the calibration

example, and selecting, using the plausibility distribution in the calibration example, the ground truth class; and determining a conformal prediction threshold for the image classification neural network using the additional calibration examples and the confidence level.

In some implementations, the method further comprises receiving a new image; processing the new image using the image classification neural network to generate a new classification distribution for the new image; generating a confidence set that includes each class that has a respective score in the new classification distribution that is greater than or equal to the conformal prediction threshold; and providing, as a conformal prediction for the new image, data identifying the confidence set.

In some implementations, for at least one of the calibration examples, the plausibility distribution includes two or more non-zero scores.

In some implementations, obtaining a respective additional calibration image from the calibration image in the calibration example comprises setting the respective additional calibration image to be the same as the calibration image.

In some implementations, selecting the ground truth class comprises: sampling, from the plausibility distribution in the calibration example, one of the plurality of classes; and selecting, as the ground truth class, the sampled class.

In some implementations, obtaining a respective additional calibration image from the calibration image in the calibration example comprises applying data augmentation to the calibration image.

In some implementations, for one or more of the calibration examples, the plausibility distribution includes only one non-zero score.

In some implementations, for each of the one or more calibration examples for which the plausibility distribution includes only one non-zero score, selecting the ground truth class comprises: selecting, as the ground truth class, the class that has the non-zero score in the plausibility distribution.

In some implementations, determining a conformal prediction threshold for the image classification neural network using the additional calibration examples and the confidence level comprises: determining, from the confidence level, a quantile value q; for each additional calibration example: processing the additional calibration image in the additional calibration example using the image classification neural network to generate a respective classification distribution; and identifying the respective classification score assigned to the ground truth class in the additional calibration example by the respective

classification distribution for the additional calibration example; and selecting, as the conformal prediction threshold, a q-th quantile of the scores assigned to the ground truth classes for the additional calibration examples.

In some implementations, determining the quantile value q comprises: determining the quantile value based on the confidence level, a total number of calibration examples n, and a total number of additional calibration examples m for each of the calibration examples.

In another aspect, a method includes obtaining a plurality of calibration examples for calibrating an image classification neural network for conformal prediction, wherein each calibration example comprises (i) a calibration image and (ii) a plausibility distribution for the calibration image that comprises a respective score for each of a plurality of classes, and wherein the image classification neural network has been trained to process an input image to generate as output a classification distribution that comprises a respective classification score for each of the plurality of classes; obtaining a confidence level that indicates a required confidence in conformal predictions generated by the image classification neural network; for each calibration example in a first set of the calibration examples, generating a plurality of first additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising, for each first additional calibration example: generating the respective additional calibration image from the calibration image in the calibration example, and selecting, using the plausibility distribution in the calibration example, the ground truth class; for each calibration example in a second set of the calibration examples, generating a second additional calibration example that includes a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising: obtaining the respective additional calibration image from the calibration image in the calibration example, and selecting, using the plausibility distribution in the calibration example, the ground truth class; determining, using the first and second additional calibration examples, an estimate of a continuous density function of p-values for the second set of calibration examples; and using the estimate of the continuous density function and the confidence level to perform conformal prediction for new input images using the image classification neural network.

In some implementations, using the estimate of the continuous density function and the confidence level to perform conformal prediction for new input images using the image classification neural network comprises: receiving a new image; processing the

new image using the image classification neural network to generate a new classification distribution for the new image; generating a respective initial p-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example, a respective classification score in a classification output for the first additional calibration example for the ground truth class in the first additional calibration example; generating a respective modified p-value for each class of the plurality of classes, comprising applying the estimate of the continuous density function to the respective initial p-value; generating a confidence set that includes each class that has a respective modified p-value that is greater than the confidence level; and providing, as a conformal prediction for the new image, data identifying the confidence set.

In some implementations, generating a respective modified p-value for each class of the plurality of classes further comprises: determining an upper bound of the estimate of the continuous density function to the respective initial p-value.

In some implementations, determining, using the first and second additional calibration examples, an estimate of a continuous density function of p-values for the second set of calibration examples comprises: for each additional calibration example: processing the additional calibration image in the additional calibration example using the image classification neural network to generate a respective classification distribution; and identifying the respective classification score assigned to the ground truth class in the additional calibration example by the respective classification distribution for the additional calibration example; generating a respective p-value for each of the second set of calibration set of calibration examples using (i) the respective classification score assigned to the ground truth class in the second additional calibration example by the respective classification distribution for the second additional calibration example and (ii) the respective classification scores assigned to the ground truth classes in the first additional calibration examples by the respective classification distributions for the first additional calibration examples; and determining the estimate from the respective p-values.

In some implementations, for at least one of the calibration examples in the first set, the plausibility distribution includes two or more non-zero scores.

In some implementations, obtaining a respective additional calibration image from the calibration image in the calibration example comprises setting the respective additional calibration image to be the same as the calibration image.

In some implementations, selecting the ground truth class comprises: sampling, from the plausibility distribution in the calibration example, one of the plurality of classes; and selecting, as the ground truth class, the sampled class.

In another aspect, a method includes receiving a new image; processing the new image using an image classification neural network to generate a new classification distribution for the new image that comprises a respective classification score for each of a plurality of classes; generating a respective initial p-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example of a plurality of first additional calibration examples, a respective classification score in a classification output for the first additional calibration example for a ground truth class in the first additional calibration example; generating a respective modified p-value for each class of the plurality of classes, comprising applying, to the respective initial p-value, an estimate of a continuous density function of p-values for a second set of calibration examples; generating a confidence set that includes each class that has a respective modified p-value that is greater than a confidence level that indicates a required confidence in conformal predictions generated using the image classification neural network; and providing, as a conformal prediction for the new image, data identifying the confidence set.

In some implementations, generating a respective modified p-value for each class of the plurality of classes further comprises: determining an upper bound of the estimate of the continuous density function to the respective initial p-value.

In some implementations, generating a respective initial p-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example of a plurality of first additional calibration examples, a respective classification score in a classification output for the first additional calibration example for a ground truth class in the first additional calibration example comprises: determining, for each first additional calibration example, whether the respective classification score in the classification output for the first additional calibration example for the ground truth class in the first additional calibration example is less than or equal to the respective classification score in the new classification distribution for the class.

In some implementations of any of the above aspects, the input images to the image classification neural network are medical images.

In some implementations of any of the above aspects, the plurality of classes correspond to different medical conditions.

In some implementations of any of the above aspects, the medical conditions are skin conditions.

In some implementations of any of the above aspects, the plurality of classes correspond to different levels of severity of one or more particular medical conditions.

In some implementations of any of the above aspects, the particular medical condition is a skin condition.

Other aspects include systems and computer storage media corresponding to any of the aspects above.

Particular embodiments of the subject matter described in this specification can be implemented so as to realize one or more of the following advantages.

Many application domains, especially safety-critical applications such as medical diagnostics, require reasonable uncertainty estimates for decision making and benefit from statistical performance guarantees.

Conformal prediction (CP) is a statistical framework that quantifies uncertainty rigorously by providing finite-sample, non-asymptotic performance guarantees. In particular, conformal prediction uses a set of calibration examples to calibrate an already-trained classification neural network that has been trained to perform a classification task. After calibration, conformal prediction transforms a classification output generated by the classification neural network for a given input into a confidence set that can include multiple classes for the classification task.

Conventional conformal prediction approaches assume that each calibration example has a single ground truth class label, i.e., belongs to a single one of the classes for the classification task.

However, in many real-world scenarios, the underlying labels for some or all of the calibration examples are ambiguous, i.e., assign non-zero likelihoods to multiple ones of the classes. For example, for some conformal prediction approaches, one-hot labels, i.e., labels that identify a single class, are obtained by aggregating multiple expert opinions using a voting procedure, resulting in a one-hot distribution.

However, when some of the experts disagree about the label for a given example, this ambiguity is not reflected in the one-hot distribution that is eventually generated, i.e., the one-hot distribution identifies the class with the most votes regardless of whether or not the vote was unanimous.

Other sources for ambiguity that are ignored by conventional approaches are described below.

This is problematic because calibration examples for real-world tasks often have an underlying "plausibility" distribution that assigns a significant amount of probability mass to multiple different classes, indicating that several different classes can plausibly be the correct ground truth class for a given example.

This specification describes techniques for performing conformal prediction by directly making use of ambiguous calibration examples and accounting for the ambiguity in the labels for these examples. By doing this, the described techniques significantly improve over conventional approaches on tasks where there is some disagreement (or other ambiguity) in the labels for the calibration examples that are available to the system, as is the case in many real-world tasks.

As one example, when performing a skin condition classification task using calibration examples that feature significant disagreement among expert annotators, conventional conformal prediction techniques under-cover expert annotations: calibrated for 72% coverage, they fall short by on average 10%. By explicitly accounting for the disagreement during calibration, the described techniques significantly close this gap both empirically and theoretically.

The details of one or more embodiments of the subject matter of this specification are set forth in the accompanying drawings and the description below. Other features, aspects, and advantages of the subject matter will become apparent from the description, the drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows an example neural network system.

FIG. 2 is a flow diagram of an example process for generating calibration data that includes a conformal prediction threshold.

FIG. 3 is a flow diagram of an example process for performing conformal prediction.

FIG. 4 is a flow diagram of an example process for generating calibration data that includes an estimate of a continuous density function.

FIG. 5 is a flow diagram of an example process for performing conformal prediction.

FIG. 6 is a flow diagram of an example process for determining the estimate of the continuous density function.

FIG. 7 shows an example of the performance of the described techniques.

FIG. 8 shows another example of the performance of the described techniques.

Like reference numbers and designations in the various drawings indicate like elements.

DETAILED DESCRIPTION

FIG. 1 shows an example neural network system 100. The neural network system 100 is an example of a system implemented as computer programs on one or more computers in one or more locations, in which the systems, components, and techniques described below can be implemented.

The system 100 is a system that performs a classification task on a data item 102 using a classifier neural network 110.

The system 100 can be configured to perform any of a variety of classification tasks. As used in this specification, a classification task is any task that that requires the system to generate, for an input data item 102, a classification output that includes a respective score for each of a set of multiple classes ("categories") and, optionally, to then select one or more of the categories as a "classification" for the data item using the respective scores.

That is, the classification output can identify the respective scores for the classes or can identify one or more selected categories and, optionally, the scores for the selected categories.

One example of a classification task is image classification, where the input data item is an image, i.e., the intensity values of the pixels of the image, the categories are object categories, and the task is to classify the image as depicting an object from one or more of the object categories. That is, the classification output for a given input image is a prediction of one or more object categories that are depicted in the input image.

Another example of a classification task is video classification, where the input data item is a video, i.e., the intensity values of the pixels of the video frames in the video, and the categories are, e.g., object categories, and the task is to classify the video as depicting an object from one or more of the object categories, action categories, and the task is to classify the video as depicting an action from one or more of the actions being performed, and so on.

Another example of a classification task is text classification, where the input data item is text and the task is to classify the text as belonging to one multiple categories. One example of such a task is sentiment analysis task, where the categories each correspond to different possible sentiments of the task. Another example of such a task is a reading comprehension task, where the input text includes a context passage and a question and the categories each correspond to different segments from the context passage that might be an answer to the question. Other examples of text processing tasks that can be framed as classification tasks include an entailment task, a paraphrase task, a textual similarity task, a sentiment task, a sentence completion task, a grammaticality task, and so on.

Other examples of classification tasks include speech processing tasks, where the input data item is audio data representing speech. Examples of speech processing tasks include language identification (where the categories are different possible languages for the speech), hotword identification (where the categories indicate whether one or more specific "hotwords" are spoken in the audio data), topic classification and so on.

More generally, the classification task can be an audio classification task, where the input data item is audio data and the categories can represent, e.g., different speakers captured in the audio data, different sound emitting objects generating sound in the audio data, different musical instruments generating music in the audio data, different animals making noises in the audio data, and so on.

The above examples may be considered examples of object recognition tasks or processes. An object may be a real-world object. That is, the object may be an object in the real-world and the input data item may be a representation of the real-world object. The representation may be obtained by a sensor (such as an imaging sensor, a microphone, etc.) sensing the object in the real-world.

As another example, the task can be an agent control task, where the input is an image or other sensor data characterizing the state of an environment and the categories correspond to different actions in a set of actions that can be performed by the agent in the environment. The system can then use the classification output to control the agent, i.e., by causing the agent to perform one of the actions from the set of actions in response to the observation. The agent can be, e.g., a robot, an autonomous vehicle, or other mechanical agent. As another example, the agent can be a control system for a facility, e.g., a data center or other facility.

As another example, the task can be a health prediction task, where the input is a sequence derived from electronic health record data for a patient and the categories are respective predictions that are relevant to the future health of the patient, e.g., a predicted treatment that should be prescribed to the patient, the likelihood that an adverse health event will occur to the patient, predictions related to the presence or absence (or the severity) of a given medical condition, or a predicted diagnosis for the patient.

For example, when the task is an image classification task and the input to the neural network 110, i.e., the data item 102, includes one or more images, the one or more images can include a medical image. A medical image is an image of corresponding patient, e.g., of all or a part of the body of the patient generated by a medical imaging device. In this example, the set of categories can correspond to different medical conditions and/or can correspond to different levels of severity of a particular medical condition.

For example, the medical conditions can be or can include skin conditions. Skin conditions can include any of acne, eczema, psoriasis, lentigo, melanoma, stasis dermatitis, alopecia areata, keratosis, skin tags, and so on.

As particular examples, the image can be a computer tomography (CT) image, a magnetic resonance imaging (MRI) image, an ultrasound image, an X-ray image, a mammogram image, a fluoroscopy image, a fundus image, or a positron-emission tomography (PET) image.

As another particular example, the one or more images can include an RGB image generated by a camera sensor, e.g., of a mobile device or a digital camera. That is, the medical imaging device can be a camera sensor that captures RGB images. As another example, the one or more images can include a greyscale image generated by a camera sensor that captures 'black and white' images.

As another example, the one or more images can be images captured by an autonomous vehicle or robot, e.g., can be images captured by a Lidar sensor, a radar sensor, or a camera sensor of the vehicle or robot.

The classification neural network 110 is a neural network that has been trained, e.g., by the system 100 or by another system to receive an input for a classification task and to generate, as output, a classification distribution 112 that includes a respective score, e.g., a probability, for each of the plurality of classes for the classification task.

The classification neural network 110 can generally have any appropriate architecture that maps an input data item to a set of scores that includes a respective score

for each class for the classification task. Example classification neural network architectures include convolutional neural networks, self-attention neural networks, e.g., vision Transformer neural networks, fully-connected neural networks, e.g., multi-layer perceptrons (MLPs), and recurrent neural network.

More specifically, the system 100 uses the neural network 110 to perform conformal prediction.

Performing conformal prediction refers to generating, using the classification distribution 112 for a given input, an output for the given input that includes a confidence set 114 of classes from the plurality of classes.

In particular, the confidence set 114 identifies the minimum number of classes so that the likelihood that the true class for the given input is included in the confidence set 114 is greater than or equal to a confidence that is indicated by a confidence level 120.

The confidence level 120 can be, e.g., fixed, or can be received as input from a user and can be represented as, e.g., a probability or other appropriate numerical value. The confidence level 120 can be set at least partially depending on the process that the classification neural network 110 has been trained to perform. For example, some processes (e.g. medical condition classification) may require a higher confidence level 120 than other processes (e.g. general image classification). As a particular example, a process for classifying a medical condition of a first severity may require a higher confidence level 120 than a process for classifying a medical condition of a second severity lower than the first severity, i.e., a process for classifying the medical condition of a lower severity may require a comparatively lower confidence level 120.

For example, when the confidence level 120 is represented as a probability α, for a classification task with K classes, the confidence set $C(X)$ for an input X satisfies:

$$\mathbb{P}\big(Y \in C(X)\big) \geq 1 - \alpha,$$

where Y is the (unobserved) true class for the input X and $\mathbb{P}\big(Y \in C(X)\big)$ is the probability that Y is one of the classes in $C(X)$.

Thus, the confidence set 114 can identify multiple classes from the plurality of classes and can identify variable numbers of classes for different input data items 102. That is, because different input data items 102 will have different classification distributions 112, the confidence set 114 can include different numbers of classes for different input data items 102.

In some implementations, the system 100 provides data identifying the confidence set 114 as output, e.g., for presentation to a user or for storage in association with the input 102.

In some other implementations, the system 100 can use another machine learning model to process the given input 102 and data identifying the classes in the confidence set 114 to select a final class as the final classification for the input 102.

For example, the other machine learning model can be a different, more computationally-expensive neural network that provides a more fine-grained identification of the final classification for the input 102.

As a particular example, when there are a large number of classes, it may be too computationally expensive to use the different neural network to directly perform classification. In this example, the system can first perform conformal prediction and then use the confidence set to generate a smaller input to the different neural network.

Prior to performing conformal prediction, the system 100 uses a set of calibration examples 130 to generate conformal prediction data 150 required to perform conformal prediction. The conformal prediction data 150 may be referred to as required data 150. Although the calibration examples 130 are not used to train the neural network and only to calibrate the conformal prediction data 150, the calibration examples 130 can also be referred to as training examples.

For example, the required data 150 can include a conformal prediction threshold. Conformal prediction thresholds are described in more detail below with reference to FIGS. 2 and 3.

As another example, the required data 150 can include an estimate of a continuous density function (CDF) of p-values for a subset of the calibration examples 130.

Estimates of CDFs are described in more detail below with reference to FIGS. 4-6.

Each calibration example 130 includes (i) a calibration input 132 and (ii) a plausibility distribution 134 for the calibration input 132 that includes a respective score for each of the plurality of classes.

Generally, for some or all of the calibration examples 130, the plausibility distribution 134 in the example 130 includes two or more non-zero values, i.e., reflects ambiguity about the true ground truth class for the corresponding calibration input or indicates that the calibration input can belong to two or more of the classes. The plausibility distribution 134 may quantify how plausible each class is to be the ground

truth class. The plausibility distribution 134 may be obtained based on one or more labels associated with a given input 132, e.g., determined from the one or more labels using a statistical model.

For example, the plausibility distribution 134 for a given input 132 can have been generated from labels for the given input 132 from two or more users that have some degree of disagreement, e.g., that identify a different class as the most likely true class for the given input. The plausibility distribution 134 may be obtained based on the two or more labels for the given input 132, e.g., determined from the two or more labels using a statistical model. For example, each label user can submit a label identifying a respective class, and the plausibility distribution can assign a probability to each class that is equal to the number of labels that identify the class divided by the total number of submitted classes.

As another example, the plausibility distribution 134 for a given input can have been generated from a ranked list of possible labels for the given input that does not identify the absolute likelihoods for the classes or that indicates that two or more of the labels in the ranked list have non-zero likelihoods. The plausibility distribution 134 may be obtained based on the ranked list possible labels for the given input 132, e.g., determined from the ranked list using a statistical model. For example, a set of partial rankings of classes can be aggregated into a plausibility distribution using an inverse rank normalization procedure.

As yet another example, the task can be a multi-label classification task in which any given input can belong to multiple classes and the label for the given input can identify multiple classes as the true classes for the given input. In this instance, the plausibility distribution 134 may include a non-zero score for multiple classes of the plurality of classes, indicating that each of the classes with a non-zero score may be a true class. The plausibility distribution 134 may be obtained based on the possible classes for the given input 132, e.g., determined from the label using a statistical model.

Thus, unlike conventional approaches to conformal prediction that assume that each calibration example belongs to exactly one of the classes, the system 100 performs conformal prediction using calibration examples 130 that include ambiguous labels.

Generally, to generate the required data 150, the system 100 can generate one or more additional calibration examples 140 from each calibration example 130 in the set.

Each additional calibration example 140 includes a respective additional calibration input 142 and identifies a ground truth class 144 for the additional calibration input 142.

The system 100 then uses the additional calibration examples 140 to generate the required data 150 for performing calibration prediction.

Once the data 150 has been generated, the system 100 uses the data 150 to perform conformal prediction for new inputs 102.

This is described in more detail below.

In particular, the description below describes that the data items 102 are images. More generally and as described above, however, the described techniques can be applied to any type of classification task.

FIG. 2 is a flow diagram of an example process 200 for generating calibration data that includes a conformal prediction threshold. For convenience, the process 200 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 200.

The system obtains a plurality of calibration examples for calibrating an image classification neural network for conformal prediction (step 202).

Each calibration example includes (i) a calibration image and (ii) a plausibility distribution for the calibration image that includes a respective score for each of a plurality of classes for the image classification task. As described above, the distribution is referred to as a "plausibility" distribution because the distribution can include a non-zero score for more than one of the classes, indicating that more than one class is a "plausible" class for the corresponding calibration image.

As described above, the image classification neural network has been trained, e.g. by the system or by another training system, to process an input that includes an input image, i.e., to process at least the intensity values of the pixels of the input image, to generate as output a classification distribution that includes a respective classification score for each of the plurality of classes.

The image classification neural network can have been trained using any appropriate technique, e.g., trained through supervised learning on a set of training data for the classification task or pre-trained through unsupervised learning and then trained through supervised learning on the set of training data.

The system obtains a confidence level that indicates a required confidence in

conformal predictions generated by the image classification neural network (step 204). For example, the confidence level can be expressed as a probability or other appropriate numerical value and can be obtained as input from the user or from another system.

For each calibration example, the system generates a plurality of additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image (step 206).

To generate an additional calibration example, the system obtains, e.g., generates, a respective additional calibration image from the calibration image in the calibration example.

For example, the system can set the additional calibration image to be the same as the calibration image or can apply data augmentation to the calibration image.

As such, generating a respective additional calibration image can include obtaining an additional calibration image, for example by copying the calibration image, accessing the calibration image, processing (e.g. augmenting or otherwise adjusting) the calibration image, or any other means of obtaining a respective additional calibration image.

Applying data augmentation can include generating an image that is one or more of (adversarially or randomly) perturbed or corrupted, rotated, flipped, and so on, relative to the original image. Incorporating data augmentation can improve the robustness of the conformal prediction performed by the system.

The system then selects, using the plausibility distribution in the calibration example, the ground truth class.

For example, when the plausibility distribution includes two or more non-zero scores, the system can sample, from the plausibility distribution in the calibration example, one of the plurality of classes and then select, as the ground truth class, the sampled class. Thus, when the plausibility distribution includes two or more non-zero scores, different ones of the additional calibration examples can have different ground truth classes, i.e., different ones of the classes with non-zero scores in the plausibility distribution.

When the plausibility distribution includes only one non-zero score, the system can select, as the ground truth class, the class with the non-zero score in the plausibility distribution.

The system determines a conformal prediction threshold for the image

classification neural network using the additional calibration examples and the confidence level (step 208).

The system can generally determine the conformal prediction threshold in any appropriate way, e.g., using any appropriate conformal prediction technique.

As one example, the system can determine, from the confidence level, a quantile value q. For example, the system can determine the quantile value based on the confidence level α, a total number of calibration examples n, and a total number of additional calibration examples m for each of the calibration examples.

In particular, the quantile value q can be equal to $\frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}$.

For each additional calibration example, the system can process the additional calibration image in the additional calibration example using the image classification neural network to generate a respective classification distribution and then identify the respective classification score assigned to the ground truth class in the additional calibration example by the respective classification distribution for the additional calibration example.

The system can then select, as the conformal prediction threshold, the q-th quantile of the scores assigned to the ground truth classes for the additional calibration examples.

As a particular example of the above, the threshold τ can satisfy:

$$Q\left(\left\{E\left(X_i, Y_i^j\right)\right\}_{i\in[n], j\in[m]}; q\right).$$

where $\left(X_i, Y_i^j\right)$ is the j-th additional calibration example generated from the i-th calibration example, Y_i^j is the ground truth class identified in the j-th additional calibration example generated from the i-th calibration example, and $E\left(X_i, Y_i^j\right)$ is the respective classification score assigned to the ground truth class Y_i^j in the additional calibration example by the respective classification distribution for the additional calibration example.

FIG. 3 is a flow diagram of an example process 300 for performing conformal prediction after determining the conformal prediction threshold. For convenience, the process 300 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 300.

The system can perform the process 300 for new inputs after performing 200 to determine the conformal prediction threshold.

The system receives a new image (step 302).

The system processes the new image using the image classification neural network to generate a new classification distribution for the new image (step 304).

The system generates a confidence set that includes each class that has a respective score in the new classification distribution that is greater than or equal to the conformal prediction threshold (step 306).

That, is, the system uses the conformal prediction threshold generated using the additional calibration examples to perform conformal prediction. Because the additional calibration examples reflect the underlying ambiguity in the plausibility distributions for the original calibration examples, the system can more accurately perform conformal prediction, i.e., more accurately generate conformal predictions that satisfy the confidence level received as input by the system.

The system provides, as a conformal prediction for the new image, data identifying the confidence set (step 308).

FIG. 4 is a flow diagram of an example process 400 for generating calibration data that includes an estimate of a continuous density function. For convenience, the process 400 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 400.

The system obtains a plurality of calibration examples for calibrating an image classification neural network for conformal prediction (step 402).

Each calibration example includes (i) a calibration image and (ii) a plausibility distribution for the calibration image that includes a respective score for each of a plurality of classes for the image classification task.

As described above, the image classification neural network has been trained, e.g. by the system or by another training system, to process an input that includes an input image to generate as output a classification distribution that includes a respective classification score for each of the plurality of classes. The image classification neural network can have been trained using any appropriate technique, e.g., trained through supervised learning on a set of training data for the classification task or pre-trained through unsupervised learning and then trained through supervised learning on the set of training data.

The system obtains a confidence level that indicates a required confidence in conformal predictions generated by the image classification neural network (step 404). For example, the confidence level can be expressed as a probability or other appropriate numerical value and can be obtained as input from the user or from another system.

For each calibration example in a first set of the calibration examples, the system generates a plurality of first additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image (step 406).

For example, the system can receive an input splitting the calibration examples into the first set of calibration examples and a second set of calibration examples. As another example, the system can randomly split the calibration examples into the first set of calibration examples and the second set of calibration examples. As another example, the system can receive an input specifying the size of the first set of calibration examples and can then randomly select calibration examples until the first set has the specified size.

The system can generate the additional calibration examples for any given calibration example in the first set as described above with reference to FIG. 2.

For each calibration example in a second set of the calibration examples, the system generates a respective second additional calibration example that includes a respective additional calibration image and a ground truth class for the respective additional calibration image (step 408).

The system can generate the additional calibration examples for any given calibration example in the second set as described above with reference to FIG. 2. However, instead of generating multiple additional calibration examples for each calibration example, the system generates only a single additional calibration example.

The system determines, using the first and second additional calibration examples, an estimate of a continuous density function (CDF) of p-values for the second set of calibration examples (step 410).

Generally, the p-value for a given test example measures the probability that the ground truth class identified in the test example is the true ground truth class for the calibration image in the test example based on (i) the respective classification score for the ground truth class in the test example in the classification output for the test example generated by the classification neural network and (ii) the respective classification scores for the ground truth classes in the calibration examples in the corresponding classification outputs.

However, because of the ambiguity reflected in the plausibility distributions, the true ground truth classes for the calibration examples are not known to the system, so the system instead uses estimates of the p-values for the second set of calibration examples when estimating the CDF.

Generally, the estimate of the p-value for a given additional calibration example measures the probability that the ground truth class identified in the additional calibration example is the true ground truth class for the calibration image in the example based on (i) the respective classification score for the ground truth class in the additional calibration example in the classification output for the additional calibration example and (ii) the respective classification scores for the ground truth classes in the first additional calibration examples in the corresponding classification outputs.

The estimate of the continuous density function is an estimate of a function that takes as input a p-value and generates as output the probability that a given test example will have a p-value less than or equal to the input p-value.

In some cases, the system also determines, as part of the estimate, an upper bound of the estimate of the continuous density function of p-values.

The system can generally determine the estimate of the continuous density function in any appropriate way, e.g., using any appropriate conformal prediction technique.

One example technique for determining the estimate is described below with reference to FIG. 6.

The system then uses the estimate of the continuous density function and the confidence level to perform conformal prediction for new input images using the image classification neural network.

FIG. 5 is a flow diagram of an example process 500 for performing conformal prediction after determining the estimate of the continuous density function. For convenience, the process 500 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 500.

The system can perform the process 500 for new inputs after performing the process 400 to determine the estimate of the continuous density function.

The system receives a new image (step 502).

The system processes the new image using the image classification neural network to generate a new classification distribution for the new image (step 504).

The system generates a respective initial *p*-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example, a respective classification score in a classification output for the first additional calibration example for the ground truth class in the first additional calibration example (step 506).

For example, for a class *k*, the system can compute the initial *p*-value as follows:

$$\bar{\rho}_k = \frac{\sum_{j=1}^{m} \left(\sum_{p=1}^{l} \mathbb{I}[E(X_p, Y_p^j) \leq E(X, k)] + 1 \right)}{m(l+1)},$$

where $\mathbb{I}$ is the indicator function and *l* is the number of calibration examples in the first set.

The system generates a respective modified *p*-value for each class of the plurality of classes (step 508).

In particular, the system applies the estimate of the continuous density function to the respective initial *p*-value. As a particular example, the system can determine the modified *p*-value as an upper bound of the estimate of the continuous density function to the respective initial *p*-value:

$$\bar{\rho}_k^{\text{corr}} = \bar{F}_+(\bar{\rho}_k).$$

The system generates a confidence set that includes each class that has a respective modified *p*-value that is greater than the confidence level (step 510).

The system provides, as a conformal prediction for the new image, data identifying the confidence set (step 512).

FIG. 6 is a flow diagram of an example process 600 for determining the estimate of the continuous density function for the second calibration examples. For convenience, the process 600 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 600.

For each additional calibration example, the system processes the additional calibration image in the additional calibration example using the image classification

neural network to generate a respective classification distribution (step 602) and identifies the respective classification score assigned to the ground truth class in the additional calibration example by the respective classification distribution for the additional calibration example (step 604).

The system then generates a respective *p*-value for each of the second set of calibration examples (step 606). As described above, because the true ground truth classes for the calibration examples may not be known to the system, the system computes estimates of the *p*-values by leveraging the split of the calibration examples into the first and second set and only computing *p*-values for the second set of calibration examples.

In particular, the system generates the respective *p*-values using (i) the respective classification score assigned to the ground truth class in the second additional calibration example by the respective classification distribution for the second additional calibration example and (ii) the respective classification scores assigned to the ground truth classes in the first additional calibration examples by the respective classification distributions for the first additional calibration examples.

As a particular example, the respective *p*-value for the *i*-th calibration example in the second set can satisfy:

$$\bar{\rho}^i = \frac{\sum_{j=1}^{m} \left(\sum_{p=1}^{l} \mathbb{I}[E(X_p, Y_p^j) \leq E(X_i, Y_i)] + 1 \right)}{m(l+1)}.$$

The system determines the estimate of the continuous density function for the second calibration examples from the respective *p*-values (step 608).

For example, the estimate can satisfy:

$$\bar{F}(f) = \frac{1}{n-l} \sum_{i=l+1}^{n} \mathbb{I}[\bar{\rho}^i \leq f]$$

As described above, the system can also determine the upper bound on the estimate. For example, the upper and lower bounds of the estimate can satisfy:

$$\bar{F}_{\pm}(f) = \frac{\min}{\max} \left\{ \bar{F}(f) \pm \sqrt{\frac{1}{2(n-l)} \log \frac{2}{\delta}}, \frac{1}{0} \right\}$$

FIG. 7 shows an example 700 of the performance of the described techniques ("Monte Carlo CP") relative to a conventional conformal prediction ("CP with voted

labels"). In particular, Monte Carlo CP refers to the described technique that uses the threshold as the required conformal prediction data.

In particular, the example 700 shows a comparison of CP with voted labels and Monte Carlo CP on two concrete examples. For each example, the example 700 shows the plausibility distributions for the corresponding input image and the confidence sets generated by both Monte Carlo CP and CP with voted labels.

Both of the examples shown in the example 700 are ambiguous cases, i.e., as shown by the high-entropy plausibility distributions that assign significant probability values to multiple different classes.

As can be seen from the example 700, Monte Carlo CP covers more plausibility mass (i.e., yields higher realized aggregated coverage) than CP with voted labels by directly making use of the plausibility distributions instead of relying on voting, potentially improving patient outcome.

FIG. 8 shows another example 800 of the performance of the described techniques.

In particular, FIG. 8 shows the performance of Monte Carlo CP and ECDF Monte Carlo CLP relative to a standard CP technique on a task that requires classifying images from a data set of images of skin conditions, i.e., a skin condition classification task. ECDF Monte Carlo CLP refers to the described techniques that use the estimate of the CDF as the required conformal prediction data.

More specifically, FIG. 8 shows a comparison between CP applied to voted labels (left), Monte Carlo CP (middle) and ECDF Monte Carlo CP (right) in terms of voted coverage and aggregated coverage (green) across random calibration/test splits (top) as well as inefficiency (bottom). Inefficiency is a measure of the size of the confidence sets generated by a given CP technique. As can be seen from the Figure, CP with voted labels does not reach the 73% aggregated target coverage while both Monte Carlo CP and ECDF Monte Carlo CP overcome this gap.

This specification uses the term "configured" in connection with systems and computer program components. For a system of one or more computers to be configured to perform particular operations or actions means that the system has installed on it software, firmware, hardware, or a combination of them that in operation cause the system to perform the operations or actions. For one or more computer programs to be configured to perform particular operations or actions means that the one or more

programs include instructions that, when executed by data processing apparatus, cause the apparatus to perform the operations or actions.

Embodiments of the subject matter and the functional operations described in this specification can be implemented in digital electronic circuitry, in tangibly-embodied computer software or firmware, in computer hardware, including the structures disclosed in this specification and their structural equivalents, or in combinations of one or more of them. Embodiments of the subject matter described in this specification can be implemented as one or more computer programs, i.e., one or more modules of computer program instructions encoded on a tangible non transitory storage medium for execution by, or to control the operation of, data processing apparatus. The computer storage medium can be a machine-readable storage device, a machine-readable storage substrate, a random or serial access memory device, or a combination of one or more of them. Alternatively or in addition, the program instructions can be encoded on an artificially generated propagated signal, e.g., a machine-generated electrical, optical, or electromagnetic signal, that is generated to encode information for transmission to suitable receiver apparatus for execution by a data processing apparatus.

The term "data processing apparatus" refers to data processing hardware and encompasses all kinds of apparatus, devices, and machines for processing data, including by way of example a programmable processor, a computer, or multiple processors or computers. The apparatus can also be, or further include, special purpose logic circuitry, e.g., an FPGA (field programmable gate array) or an ASIC (application specific integrated circuit). The apparatus can optionally include, in addition to hardware, code that creates an execution environment for computer programs, e.g., code that constitutes processor firmware, a protocol stack, a database management system, an operating system, or a combination of one or more of them. Thus a system, artificial neural network, or trained artificial neural network as described herein, can be implemented in hardware using electronic circuitry, e.g. in a physical box. Similarly computer code as described herein can be code to emulate such hardware or code for a hardware description language.

A computer program, which may also be referred to or described as a program, software, a software application, an app, a module, a software module, a script, or code, can be written in any form of programming language, including compiled or interpreted languages, or declarative or procedural languages; and it can be deployed in any form, including as a stand alone program or as a module, component, subroutine, or other unit

suitable for use in a computing environment. A program may, but need not, correspond to a file in a file system. A program can be stored in a portion of a file that holds other programs or data, e.g., one or more scripts stored in a markup language document, in a single file dedicated to the program in question, or in multiple coordinated files, e.g., files that store one or more modules, sub programs, or portions of code. A computer program can be deployed to be executed on one computer or on multiple computers that are located at one site or distributed across multiple sites and interconnected by a data communication network.

In this specification, the term "database" is used broadly to refer to any collection of data: the data does not need to be structured in any particular way, or structured at all, and it can be stored on storage devices in one or more locations. Thus, for example, the index database can include multiple collections of data, each of which may be organized and accessed differently.

Similarly, in this specification the term "engine" is used broadly to refer to a software-based system, subsystem, or process that is programmed to perform one or more specific functions. Generally, an engine will be implemented as one or more software modules or components, installed on one or more computers in one or more locations. In some cases, one or more computers will be dedicated to a particular engine; in other cases, multiple engines can be installed and running on the same computer or computers.

The processes and logic flows described in this specification can be performed by one or more programmable computers executing one or more computer programs to perform functions by operating on input data and generating output. The processes and logic flows can also be performed by special purpose logic circuitry, e.g., an FPGA or an ASIC, or by a combination of special purpose logic circuitry and one or more programmed computers.

Computers suitable for the execution of a computer program can be based on general or special purpose microprocessors or both, or any other kind of central processing unit. Generally, a central processing unit will receive instructions and data from a read only memory or a random access memory or both. The elements of a computer are a central processing unit for performing or executing instructions and one or more memory devices for storing instructions and data. The central processing unit and the memory can be supplemented by, or incorporated in, special purpose logic circuitry. Generally, a computer will also include, or be operatively coupled to receive data from or transfer data to, or both, one or more mass storage devices for storing data, e.g., magnetic,

magneto optical disks, or optical disks. However, a computer need not have such devices. Moreover, a computer can be embedded in another device, e.g., a mobile telephone, a personal digital assistant (PDA), a mobile audio or video player, a game console, a Global Positioning System (GPS) receiver, or a portable storage device, e.g., a universal serial bus (USB) flash drive, to name just a few.

Computer readable media suitable for storing computer program instructions and data include all forms of non volatile memory, media and memory devices, including by way of example semiconductor memory devices, e.g., EPROM, EEPROM, and flash memory devices; magnetic disks, e.g., internal hard disks or removable disks; magneto optical disks; and CD ROM and DVD-ROM disks.

To provide for interaction with a user, embodiments of the subject matter described in this specification can be implemented on a computer having a display device, e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor, for displaying information to the user and a keyboard and a pointing device, e.g., a mouse or a trackball, by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback, e.g., visual feedback, auditory feedback, or tactile feedback; and input from the user can be received in any form, including acoustic, speech, or tactile input. In addition, a computer can interact with a user by sending documents to and receiving documents from a device that is used by the user; for example, by sending web pages to a web browser on a user's device in response to requests received from the web browser. Also, a computer can interact with a user by sending text messages or other forms of message to a personal device, e.g., a smartphone that is running a messaging application, and receiving responsive messages from the user in return.

Data processing apparatus for implementing machine learning models can also include, for example, special-purpose hardware accelerator units for processing common and compute-intensive parts of machine learning training or production, i.e., inference, workloads.

Machine learning models can be implemented and deployed using a machine learning framework, e.g., a TensorFlow framework or a Jax framework.

Embodiments of the subject matter described in this specification can be implemented in a computing system that includes a back end component, e.g., as a data server, or that includes a middleware component, e.g., an application server, or that

includes a front end component, e.g., a client computer having a graphical user interface, a web browser, or an app through which a user can interact with an implementation of the subject matter described in this specification, or any combination of one or more such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication, e.g., a communication network. Examples of communication networks include a local area network (LAN) and a wide area network (WAN), e.g., the Internet.

The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other. In some embodiments, a server transmits data, e.g., an HTML page, to a user device, e.g., for purposes of displaying data to and receiving user input from a user interacting with the device, which acts as a client. Data generated at the user device, e.g., a result of the user interaction, can be received at the server from the device.

While this specification contains many specific implementation details, these should not be construed as limitations on the scope of any invention or on the scope of what may be claimed, but rather as descriptions of features that may be specific to particular embodiments of particular inventions. Certain features that are described in this specification in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable subcombination. Moreover, although features may be described above as acting in certain combinations and even initially be claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to a subcombination or variation of a subcombination.

Similarly, while operations are depicted in the drawings and recited in the claims in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Moreover, the separation of various system modules and components in the embodiments described above should not be understood as requiring such separation in all embodiments, and it should be

understood that the described program components and systems can generally be integrated together in a single software product or packaged into multiple software products.

Particular embodiments of the subject matter have been described. Other embodiments are within the scope of the following claims. For example, the actions recited in the claims can be performed in a different order and still achieve desirable results. As one example, the processes depicted in the accompanying figures do not necessarily require the particular order shown, or sequential order, to achieve desirable results. In some cases, multitasking and parallel processing may be advantageous.

CLAIMS

1. A method performed by one or more computers, the method comprising:

obtaining a plurality of calibration examples for calibrating an image classification neural network for conformal prediction, wherein each calibration example comprises (i) a calibration image and (ii) a plausibility distribution for the calibration image that comprises a respective score for each of a plurality of classes, and wherein the image classification neural network has been trained to process an input that comprises an input image to generate as output a classification distribution that comprises a respective classification score for each of the plurality of classes;

obtaining a confidence level that indicates a required confidence in conformal predictions generated by the image classification neural network;

for each calibration example, generating a plurality of additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising, for each additional calibration example:

obtaining the respective additional calibration image from the calibration image in the calibration example, and

selecting, using the plausibility distribution in the calibration example, the ground truth class; and

determining a conformal prediction threshold for the image classification neural network using the additional calibration examples and the confidence level.

2. The method of claim 1, further comprising:

receiving a new image;

processing the new image using the image classification neural network to generate a new classification distribution for the new image;

generating a confidence set that includes each class that has a respective score in the new classification distribution that is greater than or equal to the conformal prediction threshold; and

providing, as a conformal prediction for the new image, data identifying the confidence set.

3. The method of any preceding claim, wherein, for at least one of the calibration examples, the plausibility distribution includes two or more non-zero scores.

4. The method of claim 3, wherein obtaining a respective additional calibration image from the calibration image in the calibration example comprises setting the respective additional calibration image to be the same as the calibration image.

5. The method of claim 3 or claim 4, wherein selecting the ground truth class comprises:

sampling, from the plausibility distribution in the calibration example, one of the plurality of classes; and

selecting, as the ground truth class, the sampled class.

6. The method of preceding claim, wherein obtaining a respective additional calibration image from the calibration image in the calibration example comprises applying data augmentation to the calibration image.

7. The method of any preceding claim, wherein, for one or more of the calibration examples, the plausibility distribution includes only one non-zero score.

8. The method of claim 7, wherein, for each of the one or more calibration examples for which the plausibility distribution includes only one non-zero score, selecting the ground truth class comprises:

selecting, as the ground truth class, the class that has the non-zero score in the plausibility distribution.

9. The method of any preceding claim, wherein determining a conformal prediction threshold for the image classification neural network using the additional calibration examples and the confidence level comprises:

determining, from the confidence level, a quantile value q;

for each additional calibration example:

processing the additional calibration image in the additional calibration example using the image classification neural network to generate a respective classification distribution; and

identifying the respective classification score assigned to the ground truth class in the additional calibration example by the respective classification distribution for the additional calibration example; and

selecting, as the conformal prediction threshold, a q-th quantile of the scores assigned to the ground truth classes for the additional calibration examples.

10. The method of claim 9, wherein determining the quantile value q comprises:

determining the quantile value based on the confidence level, a total number of calibration examples n, and a total number of additional calibration examples m for each of the calibration examples.

11. A method performed by one or more computers, the method comprising:

obtaining a plurality of calibration examples for calibrating an image classification neural network for conformal prediction, wherein each calibration example comprises (i) a calibration image and (ii) a plausibility distribution for the calibration image that comprises a respective score for each of a plurality of classes, and wherein the image classification neural network has been trained to process an input image to generate as output a classification distribution that comprises a respective classification score for each of the plurality of classes;

obtaining a confidence level that indicates a required confidence in conformal predictions generated by the image classification neural network;

for each calibration example in a first set of the calibration examples, generating a plurality of first additional calibration examples that each include a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising, for each first additional calibration example:

generating the respective additional calibration image from the calibration image in the calibration example, and

selecting, using the plausibility distribution in the calibration example, the ground truth class;

for each calibration example in a second set of the calibration examples, generating a second additional calibration example that includes a respective additional calibration image and a ground truth class for the respective additional calibration image, comprising:

obtaining the respective additional calibration image from the calibration image in the calibration example, and

selecting, using the plausibility distribution in the calibration example, the ground truth class;

determining, using the first and second additional calibration examples, an estimate of a continuous density function of p-values for the second set of calibration examples; and

using the estimate of the continuous density function and the confidence level to perform conformal prediction for new input images using the image classification neural network.

12. The method of claim 11, wherein using the estimate of the continuous density function and the confidence level to perform conformal prediction for new input images using the image classification neural network comprises:

receiving a new image;

processing the new image using the image classification neural network to generate a new classification distribution for the new image;

generating a respective initial *p*-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example, a respective classification score in a classification output for the first additional calibration example for the ground truth class in the first additional calibration example;

generating a respective modified *p*-value for each class of the plurality of classes, comprising applying the estimate of the continuous density function to the respective initial *p*-value;

generating a confidence set that includes each class that has a respective modified *p*-value that is greater than the confidence level; and

providing, as a conformal prediction for the new image, data identifying the confidence set.

13. The method of claim 12, wherein generating a respective modified *p*-value for each class of the plurality of classes further comprises:

determining an upper bound of the estimate of the continuous density function to the respective initial *p*-value.

14. The method wherein determining, using the first and second additional calibration examples, an estimate of a continuous density function of *p*-values for the second set of calibration examples comprises:

for each additional calibration example:

processing the additional calibration image in the additional calibration example using the image classification neural network to generate a respective classification distribution; and

identifying the respective classification score assigned to the ground truth class in the additional calibration example by the respective classification distribution for the additional calibration example;

generating a respective *p*-value for each of the second set of calibration set of calibration examples using (i) the respective classification score assigned to the ground truth class in the second additional calibration example by the respective classification distribution for the second additional calibration example and (ii) the respective classification scores assigned to the ground truth classes in the first additional calibration examples by the respective classification distributions for the first additional calibration examples; and

determining the estimate from the respective *p*-values.

15. The method of any preceding claim, wherein, for at least one of the calibration examples in the first set, the plausibility distribution includes two or more non-zero scores.

16. The method of claim 15, wherein obtaining a respective additional calibration image from the calibration image in the calibration example comprises setting the respective additional calibration image to be the same as the calibration image.

17. The method of claim 15 or claim 16, wherein selecting the ground truth class comprises:

sampling, from the plausibility distribution in the calibration example, one of the plurality of classes; and

selecting, as the ground truth class, the sampled class.

18. A method performed by one or more computers, the method comprising:

receiving a new image;

processing the new image using an image classification neural network to generate a new classification distribution for the new image that comprises a respective classification score for each of a plurality of classes;

generating a respective initial *p*-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example of a plurality of first additional calibration examples, a respective classification score in a classification output for the first additional calibration example for a ground truth class in the first additional calibration example;

generating a respective modified *p*-value for each class of the plurality of classes, comprising applying, to the respective initial *p*-value, an estimate of a continuous density function of *p*-values for a second set of calibration examples;

generating a confidence set that includes each class that has a respective modified *p*-value that is greater than a confidence level that indicates a required confidence in conformal predictions generated using the image classification neural network; and

providing, as a conformal prediction for the new image, data identifying the confidence set.

19. The method of claim 18, wherein generating a respective modified *p*-value for each class of the plurality of classes further comprises:

determining an upper bound of the estimate of the continuous density function to the respective initial *p*-value.

20. The method of any one of claims 12-19, wherein generating a respective initial *p*-value for each class of the plurality of classes using the new classification distribution for the new image and, for each first additional calibration example of a plurality of first additional calibration examples, a respective classification score in a classification output for the first additional calibration example for a ground truth class in the first additional calibration example comprises:

determining, for each first additional calibration example, whether the respective classification score in the classification output for the first additional calibration example for the ground truth class in the first additional calibration example is less than or equal to the respective classification score in the new classification distribution for the class.

21. The method of any preceding claim, wherein the input images to the image classification neural network are medical images.

22. The method of claim 21, wherein the plurality of classes correspond to different medical conditions.

23. The method of claim 22, wherein the medical conditions are skin conditions.

24. The method of claim 21, wherein the plurality of classes correspond to different levels of severity of one or more particular medical conditions.

25. The method of claim 24, wherein the particular medical condition is a skin condition.

26. A system comprising:

one or more computers; and

one or more storage devices storing instructions that, when executed by the one or more computers, cause the one or more computers to perform the respective operations of the method of any one of claims 1-25.

27. One or more computer-readable storage media storing instructions that when executed by one or more computers cause the one or more computers to perform the respective operations of the method of any one of claims 1-25.

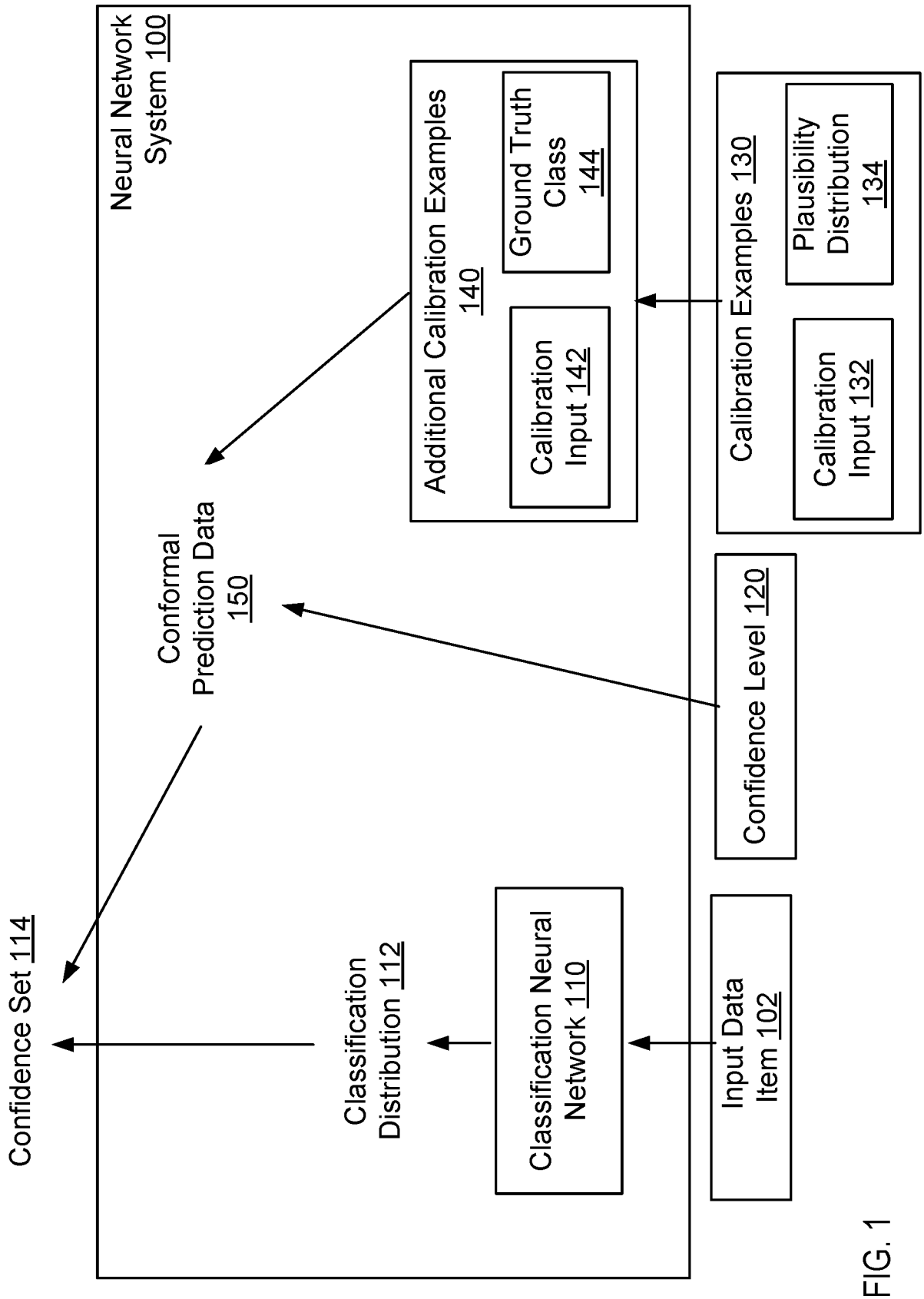

Neural Network System 100
Additional Calibration Examples 140
Ground Truth Class 144
Calibration Input 142
Calibration Examples 130
Plausibility Distribution 134
Calibration Input 132
Conformal Prediction Data 150
Confidence Level 120
Confidence Set 114
Classification Distribution 112
Classification Neural Network 110
Input Data Item 102

FIG. 1

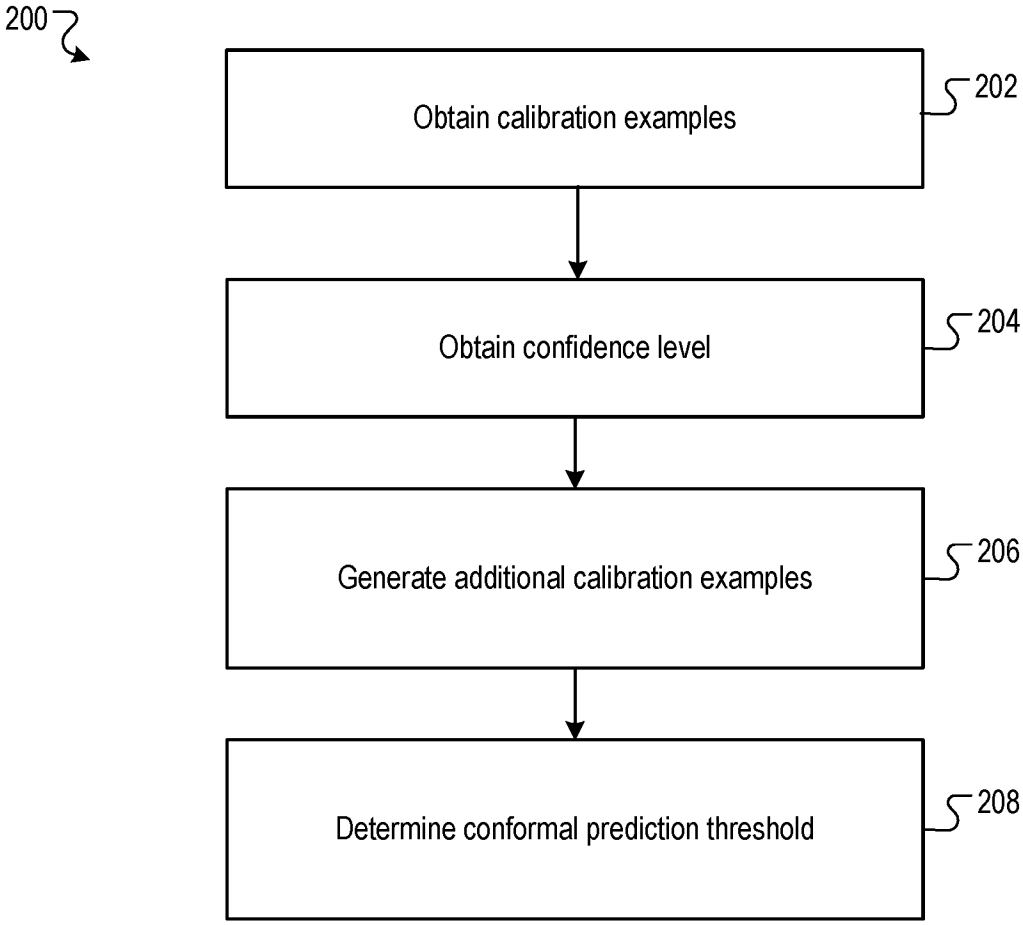

200
Obtain calibration examples
202
Obtain confidence level
204
Generate additional calibration examples
206
Determine conformal prediction threshold
208

FIG. 2

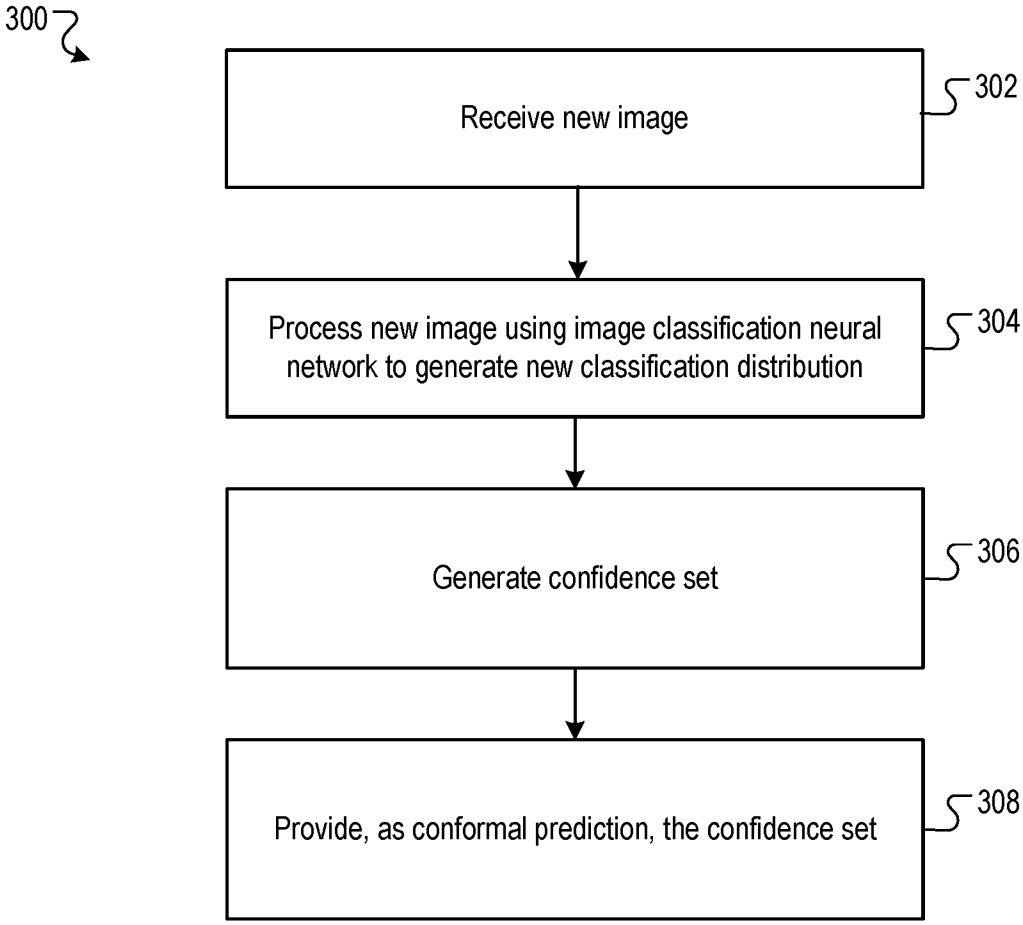

300
Receive new image
302
Process new image using image classification neural network to generate new classification distribution
304
Generate confidence set
306
Provide, as conformal prediction, the confidence set
308

FIG. 3

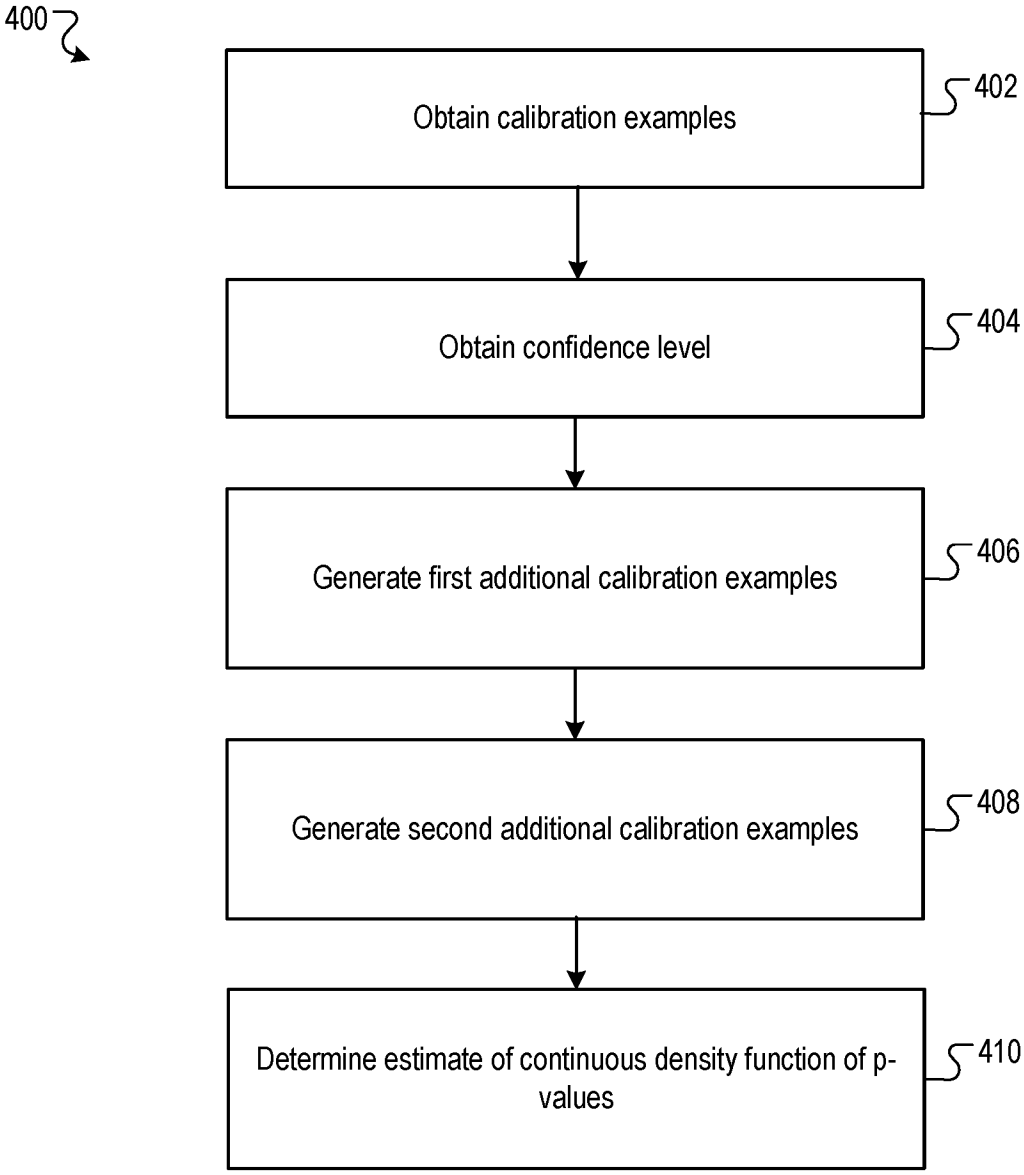

400
Obtain calibration examples
402
Obtain confidence level
404
Generate first additional calibration examples
406
Generate second additional calibration examples
408
Determine estimate of continuous density function of p-values
410

FIG. 4

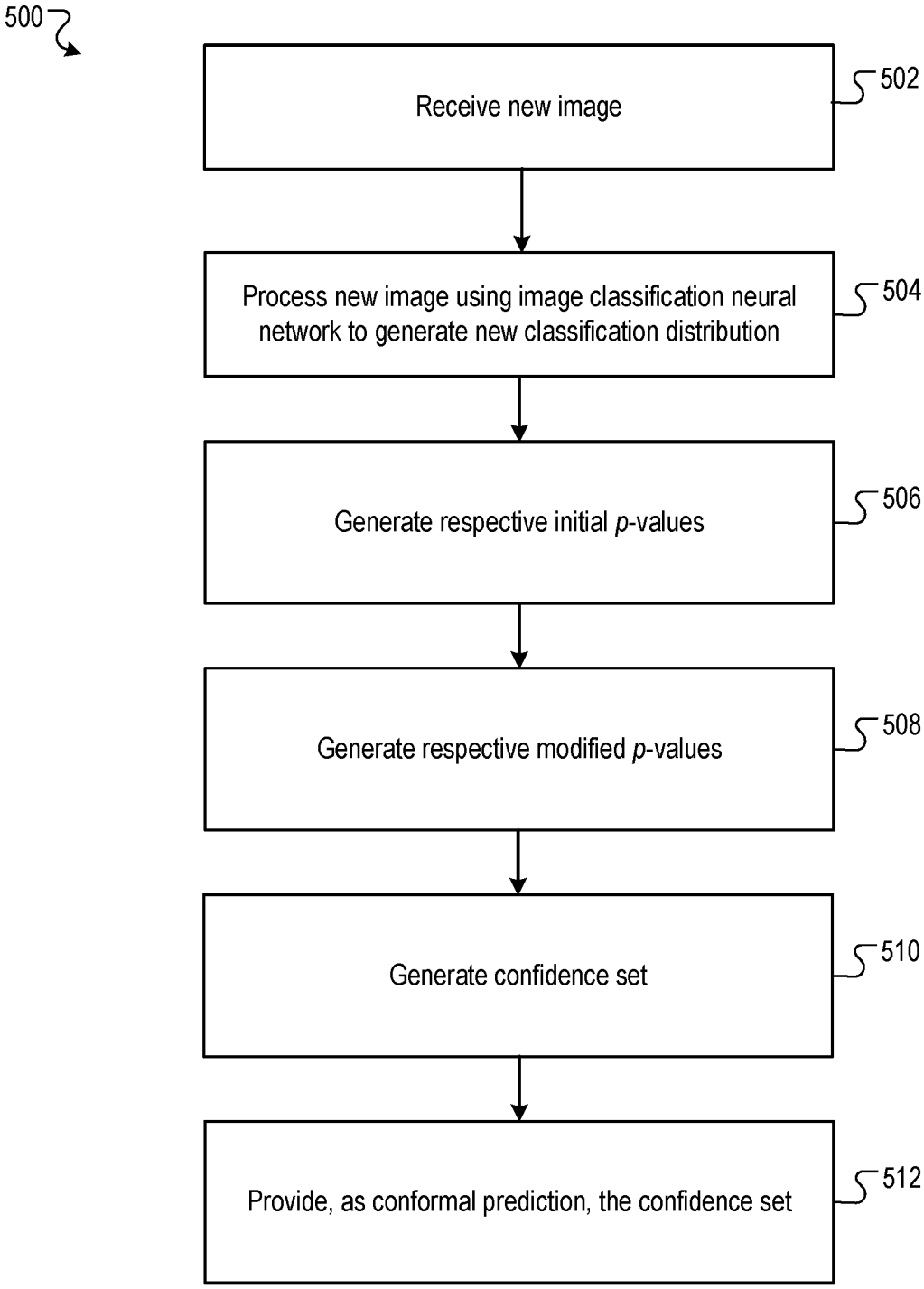

500
Receive new image
502
Process new image using image classification neural network to generate new classification distribution
504
Generate respective initial p-values
506
Generate respective modified p-values
508
Generate confidence set
510
Provide, as conformal prediction, the confidence set
512

FIG. 5

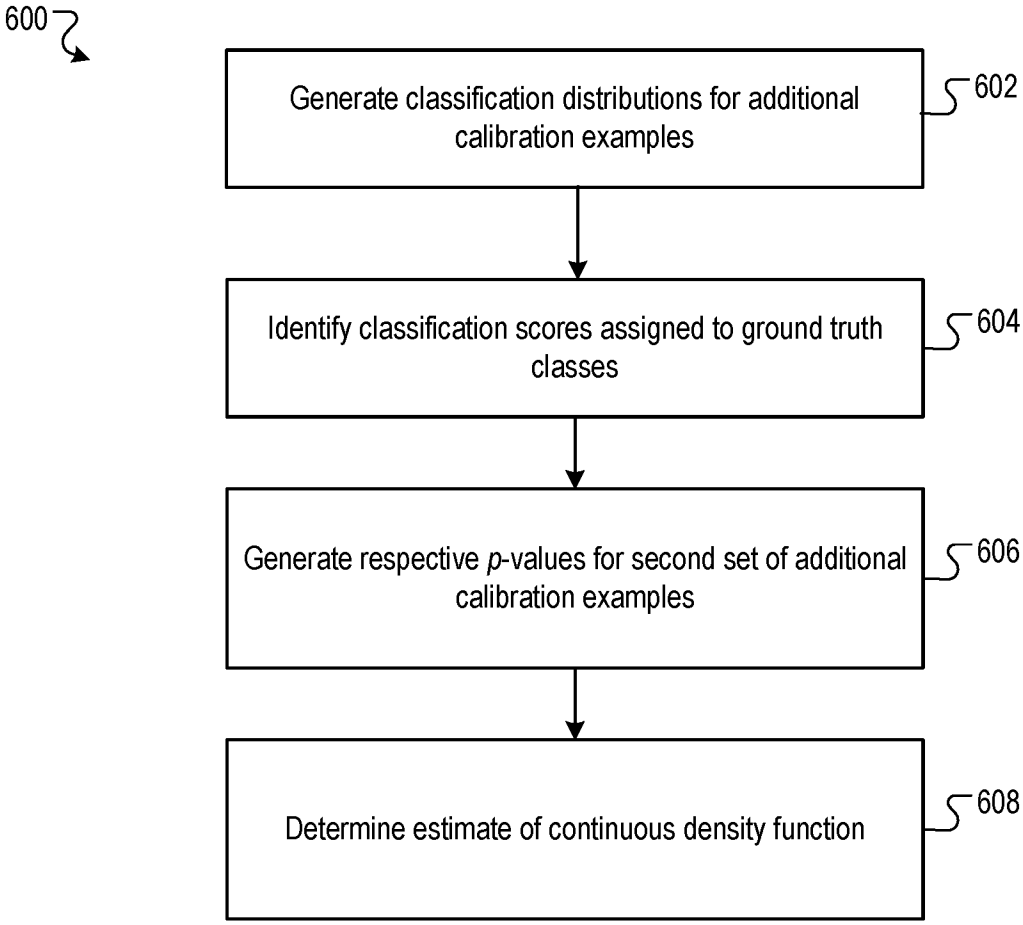

600
Generate classification distributions for additional calibration examples
602
Identify classification scores assigned to ground truth classes
604
Generate respective p-values for second set of additional calibration examples
606
Determine estimate of continuous density function
608

FIG. 6

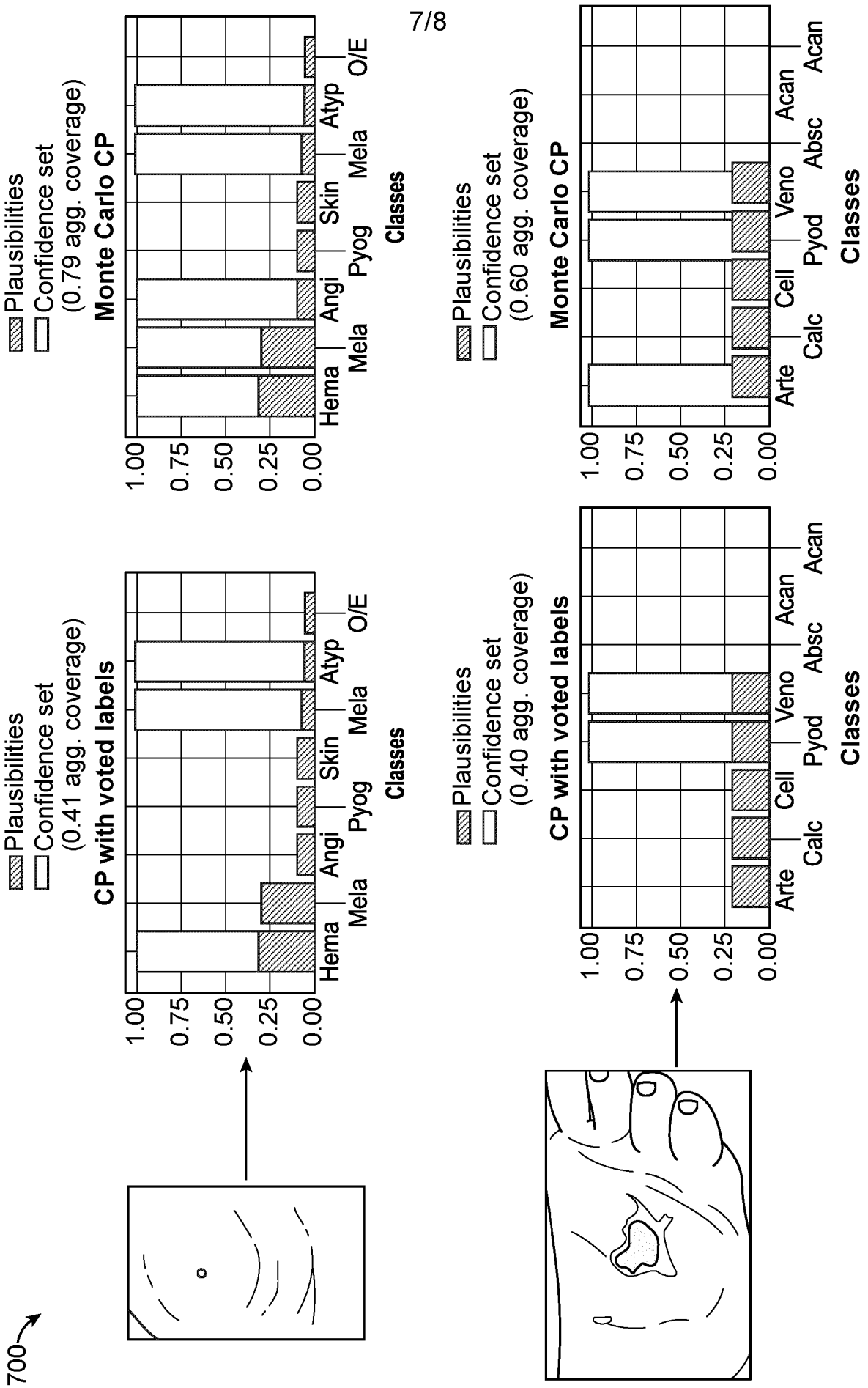

700
Plausibilities
Confidence set
(0.41 agg. coverage)
CP with voted labels
1.00
0.75
0.50
0.25
0.00
Hema
Mela
Angi
Pyog
Skin
Mela
Atyp
O/E
Classes
Plausibilities
Confidence set
(0.79 agg. coverage)
Monte Carlo CP
Plausibilities
Confidence set
(0.40 agg. coverage)
CP with voted labels
Arte
Calc
Cell
Pyod
Veno
Absc
Acan
Classes
Plausibilities
Confidence set
(0.60 agg. coverage)
Monte Carlo CP

FIG. 7

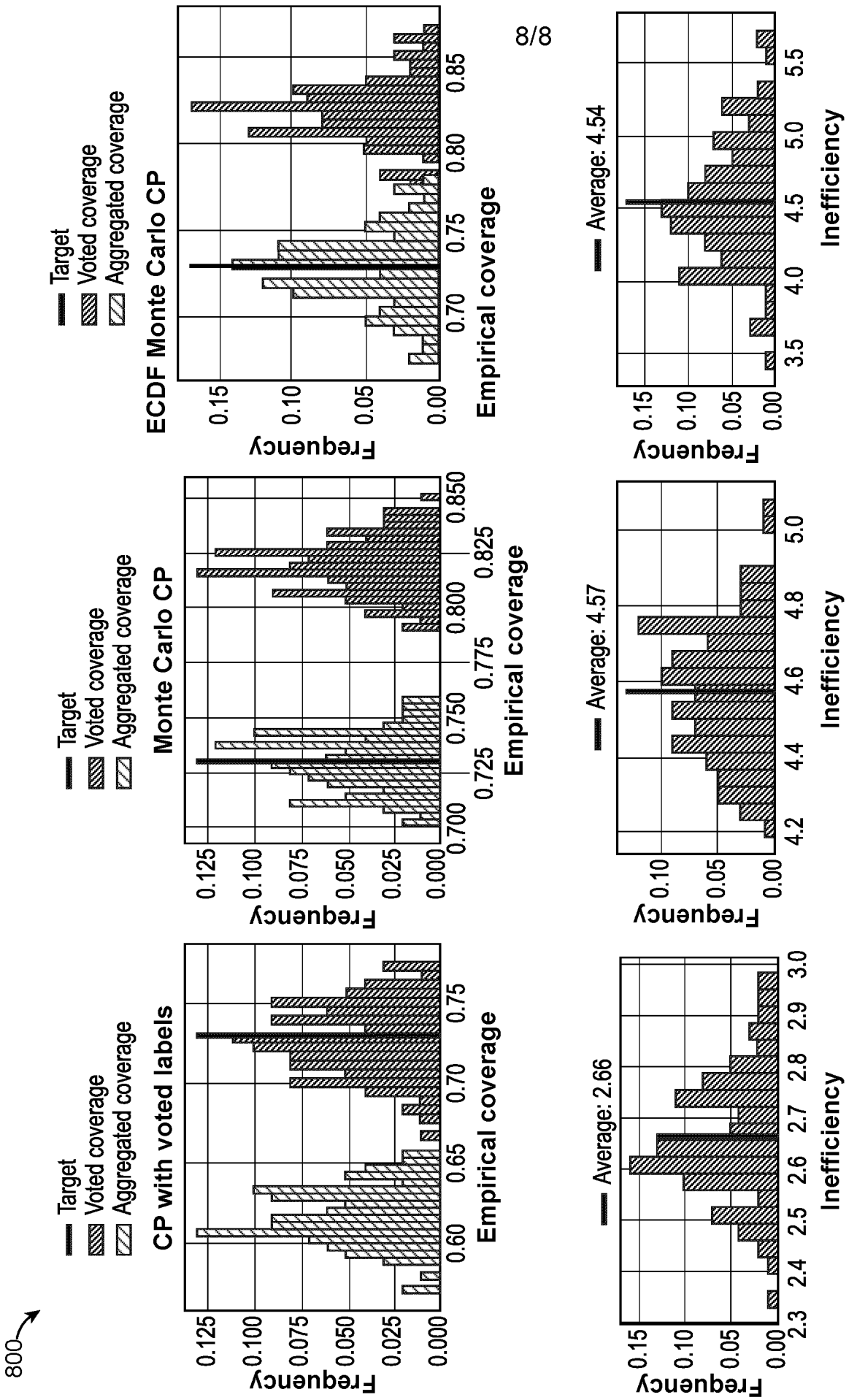

800
CP with voted labels
Monte Carlo CP
ECDF Monte Carlo CP
Target
Voted coverage
Aggregated coverage
Frequency
Empirical coverage
Average: 2.66
Average: 4.57
Average: 4.54
Inefficiency

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/070334

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06V10/764 G06N3/042 G06T7/00 G06V10/776 G06V10/82
G16H30/40 G06V10/774

ADD.

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)

G06V G06N G06T G16H

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

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

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2023/057516 A1 (DEEPMIND TECH LTD [GB]) 13 April 2023 (2023-04-13) paragraph [0002] - paragraph [0009] paragraph [0015] paragraph [0018] paragraph [0031] paragraph [0036] paragraph [0050] - paragraph [0052] paragraph [0056] - paragraph [0057] paragraph [0064] - paragraph [0065] paragraph [0091] figure 1	1-10, 21-27

-/--

[x] Further documents are listed in the continuation of Box C.

[x] 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

"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	Date of mailing of the international search report
10 October 2024	12/12/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Moreno, Marta

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DAVID STUTZ ET AL: "Evaluating AI systems under uncertain ground truth: a case study in dermatology", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 5 July 2023 (2023-07-05), XP091556012, Introduction Section 2.2 figure 9	1-10, 21-27
A	BORTOLUSSI LUCA ET AL: "Neural Predictive Monitoring", 1 October 2019 (2019-10-01), SPRINGER, Cham, Switzerland, XP047523175, ISBN: 978-3-030-32078-2 pages 129-147, figure 1 Section 3 Section 5.2	1,2,6, 26,27
A	US 2023/122168 A1 (GIBSON MOLLY KRISANN [US] ET AL) 20 April 2023 (2023-04-20) paragraph [0091]	1,6,26, 27
X,P	DAVID STUTZ ET AL: "Conformal prediction under ambiguous ground truth", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 18 July 2023 (2023-07-18), XP091566236, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853 the whole document	1-10, 21-27

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2024/070334

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims;; it is covered by claims Nos.:
1-10, 21-25(completely); 26, 27(partially)

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-10, 21-25(completely); 26, 27(partially)

   Determining a conformal prediction threshold for including labels into a set.

   ---

2. claims: 11-20(completely); 26, 27(partially)

   Providing conformal prediction for a new input image.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2024/070334

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2023057516 A1	13-04-2023	JP 2024538690 A	23-10-2024
		KR 20240055790 A	29-04-2024
		US 2024394541 A1	28-11-2024
		WO 2023057516 A1	13-04-2023
US 2023122168 A1	20-04-2023	CA 3165655 A1	05-08-2021
		CN 115668383 A	31-01-2023
		EP 4097725 A1	07-12-2022
		IL 295001 A	01-09-2022
		JP 2023512066 A	23-03-2023
		KR 20230018358 A	07-02-2023
		US 2023122168 A1	20-04-2023
		WO 2021155245 A1	05-08-2021