Received: 14 July 2021 | Revised: 7 December 2021 | Accepted: 10 December 2021

DOI: 10.1002/esp.5305

RESEARCH ARTICLE

ESPL WILEY

Comparing geomorphological maps made manually and by deep learning

W. Marijn van der Meij[1,2] | Erik W. Meijles[3] | Diego Marcos[4] | Tom T. L. Harkema[1] | Jasper H. J. Candel[5] | Gilbert J. Maas[1]

[1]Soil, Water and Land Use Team, Wageningen Environmental Research, Wageningen, The Netherlands

[2]Institute of Geography, University of Cologne, Cologne, Germany

[3]Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands

[4]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, The Netherlands

[5]Soil Geography and Landscape Group, Wageningen University & Research, Wageningen, The Netherlands

Correspondence
W. Marijn van der Meij, Soil, Water and Land Use Team, Wageningen Environmental Research, P.O. Box 47, 6700AA Wageningen, The Netherlands.
Email: m.vandermeij@uni-koeln.de

Funding information
Ministerie van Landbouw, Natuur en Voedselkwaliteit; Dutch Ministry of Agriculture, Nature and Food Quality (LNV)

Abstract

Geomorphological maps provide information on the relief, genesis and shape of the earth's surface and are widely used in sustainable spatial developments. The quality of geomorphological maps is however rarely assessed or reported, which limits their applicability. Moreover, older geomorphological maps often do not meet current quality requirements and require updating. This updating is time-consuming and because of its qualitative nature difficult to reproduce, but can be supported by novel computational methods. In this paper, we address these issues by (1) quantifying the uncertainty associated with manual geomorphological mapping, (2) exploring the use of convolutional neural networks (CNNs) for semi-automated geomorphological mapping and (3) testing the sensitivity of CNNs to uncertainties in manually created evaluation data.

We selected a test area in the Dutch push-moraine district with a pronounced relief and a high variety of landforms. For this test area we developed five manually created geomorphological maps and 27 automatically created landform maps using CNNs. The resulting manual maps are similar on a regional level. We could identify the causes of disagreement between the maps on a local level, which often related to differences in mapping experience, choices in delineation and different interpretations of the legend. Coordination of mapping efforts and field validation are necessary to create accurate and precise maps. CNNs perform well in identifying landforms and geomorphological units, but fail at correct delineation. The human geomorphologist remains necessary to correct the delineation and classification of the computed maps. The uncertainty in the manually created data that are used to train and evaluate CNNs have a large effect on the model performance and evaluation. This also advocates for coordinated mapping efforts to ensure the quality of manually created training and test data. Further model development and data processing are required before CNNs can act as standalone mapping techniques.

KEYWORDS
convolutional neural network, deep learning, geomorphological map, geomorphology, mapping, The Netherlands, uncertainty

1 | INTRODUCTION

Geomorphological maps provide information on the relief, landforms and genesis of the earth's surface, and have been developed all around the world (Bishop et al., 2012; Paron & Claessens, 2011; Verstappen, 2011). Their use is not only of scientific significance, but is also important for all kinds of spatial developments at the regional scale, such as natural hazard mapping (Chelli et al., 2021), urban development

(Douglas, 2020), archaeological prospecting (Van Lanen et al., 2015), land-use planning (Bocco et al., 2001; Roccati et al., 2020) and climate adaptation (Wageningen University and Research, 2020). However, geomorphological mapping is time-consuming and expensive, done by well-trained geomorphologists, and requires high time investment. In addition, geomorphological maps are difficult to reproduce because of interpretation and delineation differences between individuals. Hence, the quality of geomorphological maps depends on their skills, knowledge, experience and readily available data (e.g. Aspinall & Pearson, 1995; Lark et al., 2014; Randle et al., 2018). There is a need for more analysis on the background of interpretation differences between geomorphologists and ways of quantifying agreement and dissimilarities between the produced maps.

The increasing open availability and standardization of spatial data comes with an increasing demand for quality assessments of this spatial data (European Commission, 2021). Also, at a national level, quality assessment of spatial data, such as geomorphological and soil maps, has become more prominent. For example, in The Netherlands, the geomorphological map has become part of the Dutch National Key Registry of the Subsurface (*Basisregistratie Ondergrond*, BRO in Dutch), which is a central database that contains all public data and models of the Dutch subsurface (BRO, 2021). The Key Registry requires an estimate of the quality from its entries. Such quality assessments help the makers to guarantee, quantify and communicate the quality of their products and help users appropriately use the data.

In addition to a more stringent quality control, recent advances in spatial data collection, software development and computational technology and power enable the use of more advanced computational and statistical methods to aid geomorphological mapping (Bishop et al., 2012; Dramis et al., 2011; Giaccone et al., 2021). Such semi-automated mapping techniques use explanatory map layers, such as, for example, digital elevation models (DEMs) and their derivatives (also called land-surface parameters, LSPs) and maps derived from remote sensing, to identify and map geomorphological objects. Semi-automated mapping techniques can be roughly grouped into pixel-based approaches and object-based approaches. Pixel-based approaches assign a class to each pixel of a raster map, based on the information that is present for that same pixel on all explanatory maps. Object-based approaches extend this classification by also taking into account spatial context and information, such as typical sizes and shapes of the objects, representative positions for the objects and hierarchical orders of landforms (Anders et al., 2011; Drăguţ & Blaschke, 2006; Dramis et al., 2011; Feizizadeh et al., 2021; Kassouk et al., 2014). Both techniques make use of machine learning techniques for classifying pixels of a raster into a geomorphological unit (Giaccone et al., 2021; Ma et al., 2017; Valentine & Kalnins, 2016). Machine learning makes use of self-learning computer algorithms and statistical techniques to perform tasks like classification and prediction, often using large amounts of human-annotated data. Different machine learning techniques have been used in a large variety of environmental analyses. Examples are random forests in digital soil mapping (Sekulić et al., 2020), support vector machines (SVMs) in earthquake-triggered landslide susceptibility (Xu et al., 2012), groundwater spring mapping with artificial neural networks (Corsini et al., 2009) and convolutional neural networks (CNNs) in land use and land cover mapping (Zhu et al., 2017) and digital soil mapping (Wadoux et al., 2020).

A promising machine learning technique for geomorphological mapping is deep learning, particularly CNNs. CNNs are algorithms designed for pattern recognition on spatial or temporal data. These models learn from examples how to generalize the data to detect the desired patterns, eliminating the need for variable selection or feature design. CNNs can be used for a variety of tasks, such as speech recognition (Deng et al., 2013), text analysis (Young et al., 2018), image classification (Krizhevsky et al., 2017) and semantic image segmentation (Ronneberger et al., 2015). Particularly semantic segmentation (i.e. classifying each pixel of an image into a single semantic class) can be useful for geomorphological mapping. Due to the combination of spatial data generalization and segmentation, CNNs can be seen as a combination of object-based and pixel-based mapping techniques. This promising method is increasingly applied for mapping landforms and geomorphology (Abolt & Young, 2020; Bhuiyan et al., 2020; Du et al., 2019; Li et al., 2020; Palafox et al., 2017; Shumack et al., 2020; Verschoof-Van der Vaart & Lambers, 2019).

Taking these developments into consideration, the process of geomorphological mapping could be made more time-efficient and reproducible. On top of that, the quality should be properly described. Therefore, in this research we aim to (1) quantify the uncertainty associated with manual geomorphological mapping, (2) explore the use of CNNs for semi-automated geomorphological mapping and (3) test the sensitivity of CNNs to uncertainties in the manually created evaluation data.

To achieve our goals, we compare the manual mapping results of five different geomorphologists in a case study setting and carry out a semi-automated mapping procedure in the same area. The central push-moraine district in The Netherlands offers a suitable case-study area, considering its distinct landforms and wide availability of supporting, high-resolution DEMs and subsurface data and rich mapping history (Pierik & Cohen, 2020).

2 | DUTCH GEOMORPHOLOGICAL MAP AND STUDY AREA

2.1 | Legend of the geomorphological map of The Netherlands

The geomorphological map of The Netherlands (scale 1:50 000) is a nationwide morphogenetical map. The map classification contains a distinction between geomorphological units based on both their morphometry as well as their genesis. This genetical approach is essential to distinguish geomorphological units in the flatter parts of The Netherlands, where differences in relief can be minimal. The geomorphological map of The Netherlands was mostly developed in the 1960s–1990s through field surveys, soil corings and low-resolution elevation measurements (Koomen & Maas, 2004). The map is currently being updated using detailed digital elevation data and high-resolution subsurface information (Van der Meij & Maas, 2020). The map is part of the Key Registry Subsurface in The Netherlands. Wageningen Environmental Research is responsible for the maintenance of this map.

The geomorphological map is a digital polygon vector file that can be visualized and analysed in a geographic information system (GIS). Each polygon in the map is described by several legend units, which can be separately visualized. The legend of the Dutch

geomorphological map has a hierarchic structure (Maas et al., 2021). The different legend units describe the landform, genesis, relief and optional additions of a geomorphological unit (Figure 1). At the upper level, the landform describes the morphology of a landform, such as ridges, terraces, plains or valleys. At the second hierarchical level, the genesis describes which type of processes created the landform, for example periglacial, aeolian, marine or anthropogenic processes. The combination of a landform and genesis unit, together with a unique index number, forms the geomorphological unit. In the example of Figure 1, the landform R (valley), with genesis 2 (periglacial) and index number 1 forms the geomorphological unit 'dry valley'. At the third level, a relief unit and optional additional units can be added to describe relief, atypical surface sediment covers and optional active geomorphological processes. The legend of the geomorphological map of The Netherlands (in Dutch) can be found at https://legendageomorfologie.wur.nl/. This online legend contains definitions and descriptions of each legend entry.

In this study, we evaluate the manual geomorphological maps at the legend level of geomorphological units (Figure 1). The geomorphological units describe both the shape and genesis of the various landforms; most of the names of the geomorphological units are more widely used in geomorphological research and literature and this legend level is the standard level of visualization for the geomorphological map.

2.2 | Study area

The study area of Utrechtse Heuvelrug is located in the central push-moraine district of The Netherlands. We have chosen this area for its distinct relief, the occurrence of large regional landforms in combination with small local landforms, which are both relatively easily distinguishable on DEMs and in the field. The geomorphological history and shallow subsurface geology of this area are well understood and documented (Jongmans et al., 2015; Maarleveld & Van der Schans, 1961; Stouthamer et al., 2020; Van den Berg & Beets, 1987; Zagwijn, 1974). Also, a large amount of reliable corings are readily available (BRO, 2021; Van der Meulen et al., 2013). The geomorphological map of the Utrechtse Heuvelrug was first mapped in 1975 (Ten Cate et al., 1975) and was most recently updated in 2019. This provided enough data for individual mapping by geomorphologists and training the CNNs to learn how to automatically recognize the geomorphological units.

The Utrechtse Heuvelrug is a push moraine that was formed in the Saalian ice age and is currently located around 50 m higher than its surroundings. At the end of the Saalian, melting ice formed large outwash plains and slopes in front of the push moraine, on the western and southern side (Van den Berg & Beets, 1987). In the Weichselian, running water from melted snow was prevented from infiltration due to permafrost and as a consequence created incised valleys in the push moraine. These landforms are currently still present as dry valleys. Aeolian processes during the same period deposited a thick layer of coversands over large parts of the area. These coversands can occur in belt-like landforms around the push moraines, especially towards the east due to the prevailing western winds, sometimes referred to as 'belted coversands' (Maarleveld & Van der Schans, 1961). In more recent times, during the Middle Ages, parts of these coversands got remobilized by agricultural (over)exploitation of the area, forming large drift sand complexes with steep dunes and accompanying plains and depressions (Pierik et al., 2018). Also, mining activity left its mark on the landscape, in the form of deep quarries that are currently filled with water. Currently, the area is mostly covered by forest and there are no large urban areas. The most commonly occurring geomorphological units in the area, according to the legend of the geomorphological map (Maas et al., 2021), are described in Figure 2. More detailed descriptions of the legend entries are provided in the Table S1. We are mainly interested in natural geomorphological units and ignore anthropogenic structures such as ditches and roads. Anthropogenic structures that change the landscape functions, such as deep quarries and large excavations, are included in the map.

3 | METHODS

In this section we describe how the different geomorphological maps in this study were produced and evaluated. Because of the large number of different maps used and created in this study, we clarify their names and sources in Table 1.

3.1 | Available data

For the research, two main datasets were available. The first dataset is the *2019 geomorphological map* of The Netherlands (scale 1:50 000;

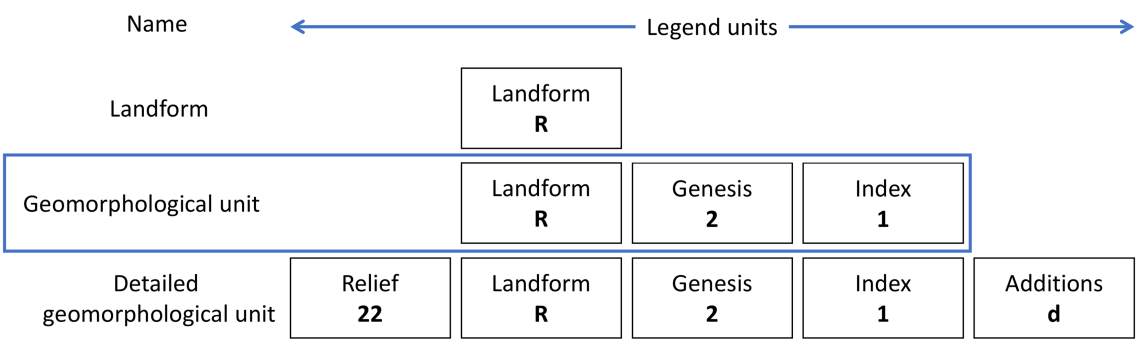

FIGURE 1 Structure of the legend of the geomorphological map of the Netherlands. The blue box indicates the legend level that is used to evaluate the manual geomorphological maps [Color figure can be viewed at wileyonlinelibrary.com]

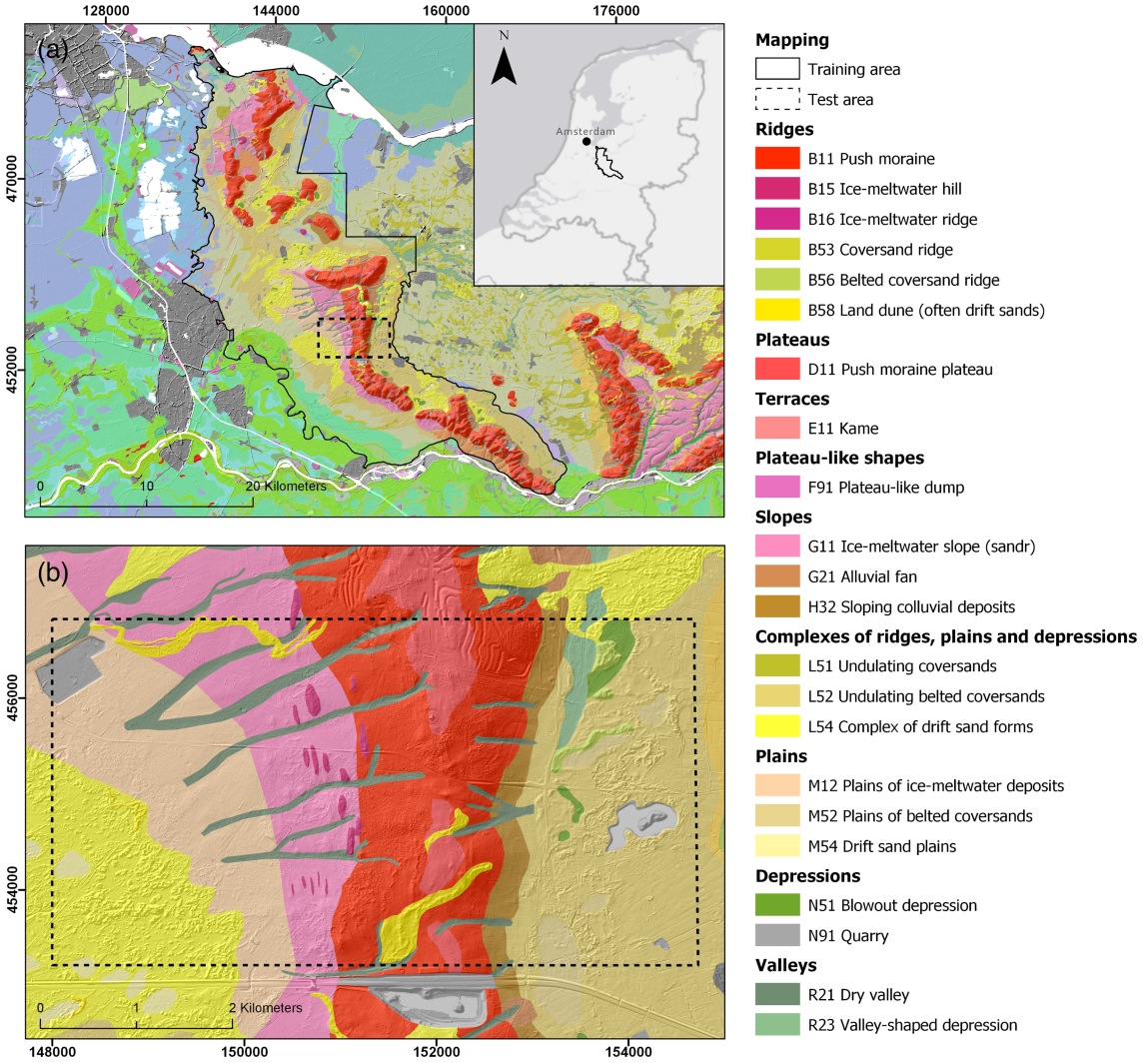

FIGURE 2 The geomorphological map of the study area Utrechtse Heuvelrug. Map A shows the training area, which has been used to train the CNNs. Map B shows the test area, for which the manual maps and computed maps were developed and evaluated. The displayed maps are from the 2019 geomorphological map (Table 1). This specific area was updated in 2019 and adjusted with the results from this study. The base map for both maps is a hillshade map of the DEM. The reference system used is RD New (https://epsg.io/28992) [Color figure can be viewed at wileyonlinelibrary.com]

Table 1). The structure of the map and its legend have been explained. The geomorphological map of the study area Utrechtse Heuvelrug was updated in 2019 with a desk study and digital information. The findings from this study have been incorporated into the update, which will be published in the next version of the Dutch geomorphological map. The 2019 geomorphological map was divided into a training area (550 km^2) and a test area (24 km^2; Figure 2). The map of the test area was not provided to the manual mappers or to the CNN in the training phase.

The second dataset is a DEM, for which we used the *AHN2* (Actueel Hoogtebestand Nederland version 2, https://www.ahn.nl/). The AHN2 is a LiDAR-derived DEM that was published in 2012. It has a cell size of 5 m. We used the AHN2 instead of the newer AHN3, because we had a version of AHN2 available where all no-data values were filled. Since the study area considers a rural area, the difference with AHN3 is expected to be minimal. The DEM shows a detailed representation of the Dutch land surface, including anthropogenic structures like roads, trails and ditches, but without buildings and vegetation.

3.2 | Manual mapping

To quantify the uncertainty associated with manual geomorphological mapping, we needed several maps of the test area, independently created by different geomorphologists. For this, we approached nine experts working at different universities and institutes in The Netherlands, from which a team of five geomorphologists was formed with different backgrounds and experience ranging from 1 to 40 years. The geomorphologists were familiar with the area, but have not worked there professionally. Three team members work at Wageningen Environmental Research. The two others work at Wageningen University and the University of Groningen. Four of the geomorphologists followed the same educational Master's programme at Wageningen University, although graduating in different years.

The mapping process was initiated with an introductory workshop, where the principles of the geomorphological map, its legend and the available data were presented, and the mapping protocol (Van der Meij & Maas, 2020) was discussed. Each team member subsequently received a data package to use during the mapping, consisting

TABLE 1 Overview of the different types of maps used, made and described in this study

Map name	Description
2019 geomorphological map	Geomorphological map of the Utrechtse Heuvelrug, updated in 2019 using manual mapping and which served as a basis for the different analyses in this study.
Manual maps	Geomorphological maps of the test area created by manual mapping (M1–M5).
Composite map	Map of the test area created by overlaying all five manual maps and determining the most commonly mapped geomorphological unit for each raster cell.
Agreement map	Map accompanying the composite map, showing the agreement between the manual maps.
Computed maps	Landform maps of the test area created by semi-automated mapping with a CNN.
Consensus map	Final geomorphological map of the test area, created by processing all findings from this study into a single map. We consider this map as the most representative map of the test area.

of the DEM and hillshade map of the entire Utrechtse Heuvelrug, and the 2019 geomorphological map of the training area, excluding the test area. Everyone was free to use other data to assist during the modelling. Other data that were used are, for example, shallow geological models (Stafleu & Dubelaar, 2016) and historical land-use maps, which are freely available for The Netherlands. Each team member independently created a geomorphological map of the test area as a desktop exercise, resulting in five geomorphological maps, named *manual maps* in this study (Table 1). One of these manual maps was the map that was made for the 2019 geomorphological map. The mapping was performed in ArcMap. There was no discussion about mapping methods and classification decisions between the team members during the creation of the manual maps. For efficiency reasons, it was expected that each geomorphologist would not spend more than 24 h on the mapping. There were no guidelines provided for mapping scale. Although the application scale of the geomorphological map of The Netherlands is legally set to 1:50 000, most mappers mapped at a more detailed scale (~1:20 000).

After the mapping, the team jointly carried out a field survey to discuss their created maps and validate the areas with most discrepancies using qualitative field observations and soil corings. The findings from this field validation were processed into the most detailed manual map, creating the *consensus map* (Table 1), which was jointly considered by the team as the most representative geomorphological map of the study area. The findings from the semi-automated mapping procedure were not included in this map. The manual maps and their quality, presented in this study, are incorporated into the quality document of the Dutch geomorphological map as well (Van der Meij & Maas, 2020).

3.3 | Semi-automated mapping

CNNs are models that can learn to recognize patterns or objects by applying a series of mathematical operations on the input data (Albawi et al., 2017). We used a CNN to recognize geomorphological patterns on images, in our case a DEM and a relief map.

The mathematical operations comprise convolutions, max pooling operators, batch normalizations and rectified linear unit activation functions (ReLU), which will be explained here respectively. Convolutions are matrix calculations with moving-window filters (kernels), typically of 3×3 cells, that are used to summarize local spatial information in the input images. This operation is similar to moving-window calculations for determining slope or topographic properties from a DEM. Pooling operators are methods that reduce the extent of the input images by locally aggregating them, in our case by passing on the maximum value in the aggregation of 2×2 cells. This pooling will increase the spatial dimensions, or receptive field, over which the convolutions can summarize spatial information, but this comes with a loss of spatial information due to the aggregation. Batch normalizations are functions that normalize the output of the convolutions, to make the model faster and more stable. A ReLU is a function that is designed to introduce non-linearity in the calculations of the CNN, by setting the negative outputs of a convolutional calculation to zero, while the positive values remain the same. CNNs need this non-linearity in order to be able to fit non-linear functions. Otherwise, applying multiple linear operations, such as convolutions, one after the other would only allow the CNN to learn a linear function of the input, which would not be suitable to solve most tasks.

The spatial dimensions of the input images are reduced due to the pooling operators. This means that spatial information gets lost. Although this may be a desirable behaviour for image classification, where the image is assigned to one single class, we would like to preserve the spatial information contained in the input images in order to classify each pixel into its corresponding geomorphological unit. For this reason, we selected U-Net (Ronneberger et al., 2015), a CNN architecture designed for semantic segmentation, or pixel-wise classification, in which the aim is to obtain a semantic map in which each individual pixel in the image is assigned to a class. In the U-Net model, the loss in spatial information is compensated by upsampling deeper layers in the model and combining them with shallower ones where the spatial information is richer. This architecture allows the model to obtain a larger receptive field, while the spatial detail is maintained. The combination of a larger receptive field and preservation of spatial information allows for semantic segmentation. Moreover, this combination enables associating each pixel in the input images with a certain class, instead of giving one class to the entire image. Semantic segmentation CNNs such as U-Net can thus recognize objects based on local information, while considering the larger context (Ronneberger et al., 2015). This makes them suitable for geomorphological mapping.

We used a symmetric U-Net architecture with four encoding layers, using two sets of 16, 32, 64 and 128 convolutional filters, respectively, all followed by a batch normalization operation and a

ReLU activation, and the same number of convolution-transposed filters in the upsampling network (Figure 3). All kernels have a 3 × 3 spatial footprint. The max pooling layers and convolution transpose layers aggregate or de-aggregate the images with 2 × 2 cells. The U-Net was trained using the stochastic gradient descent algorithm, to find the model parameters with the best fit between the predicted and actual classes. This fit was calculated using a multi-class cross-entropy loss function using a log-softmax activation on the output scores. The model trained for 5000 iterations (epochs). The initial learning rate was 0.001 and decayed following an inverse proportional law with respect to the epoch number.

U-Net requires two types of data sources for training: ground-truth example maps with objects that need to be predicted and input images on which to recognize these objects. All the data was provided in raster format. The ground-truth objects were derived from the 2019 geomorphological map for the training area, excluding the test area. This area was used for the evaluation of the results.

In order to be trained well, U-Net requires a limited amount of input classes which are sufficiently present in the training area. To satisfy this requirement, we created three subsets from the 2019 geomorphological map (Table 2). With these three subsets, we wanted to test how U-Net performed in mapping different types of landforms or geomorphological units. Subset 1 contains the main landforms in the study area, representing the highest level of the legend of the geomorphological map (Figure 1). We made a distinction between the push moraine and other types of ridges, which are often located on top of the push moraine. The aim of this subset was to create a spatially exhaustive landform map. In subset 2 we selected the dry valleys, as a representative geomorphological unit for distinct local geomorphological units. The areas that were not a dry valley on the 2019 geomorphological map were classified as a background class. In subset 3 we selected all drift-sand-related geomorphological units. These units are representative for distinct regional geomorphological units. Also here, the parts of the 2019 geomorphological map that were not related to drift sand were classified as a background class. For each subset, the 2019 geomorphological map for the training area was converted to a raster map, where each polygon was converted to a certain class following the selection criteria from Table 2. It should be noted that the mapping tasks for U-Net were different from the manual mappers. The manual mappers created a complete geomorphological map of the test area, while U-Net only mapped the landforms or certain geomorphological units in the test area. The number of classes in a complete geomorphological map would have been too much for U-Net to be trained well. Next to that, U-Net would not have been able to distinguish similar geomorphological units with a different genesis, because U-Net only had elevation data as input (see next paragraph), while other data about substrate is necessary to determine the correct genesis.

The classes in each subset are unevenly distributed in the study area. To avoid a bias in the model training towards the more commonly occurring classes, we assigned weights to each class during the training of the model. These weights were based on the relative surface area of each class in the training area. We calculated three sets of weights for each subset of geomorphological units, by taking the square root, the absolute values and the square of the inverse of the relative surface areas of each class.

The images on which U-Net should recognize the classes from each subset were derived from the DEM. The DEM contained a lot of anthropogenic structures, such as roads and ditches, which could affect the training and predictions of U-Net. Therefore, we created a smoothed digital terrain model (DTM) using a Gaussian filter of 5 × 5 cells, largely to remove these anthropogenic structures. This filtering removed a large part of the anthropogenic structures, but also some characteristics of the target objects, such as small dunes, which are

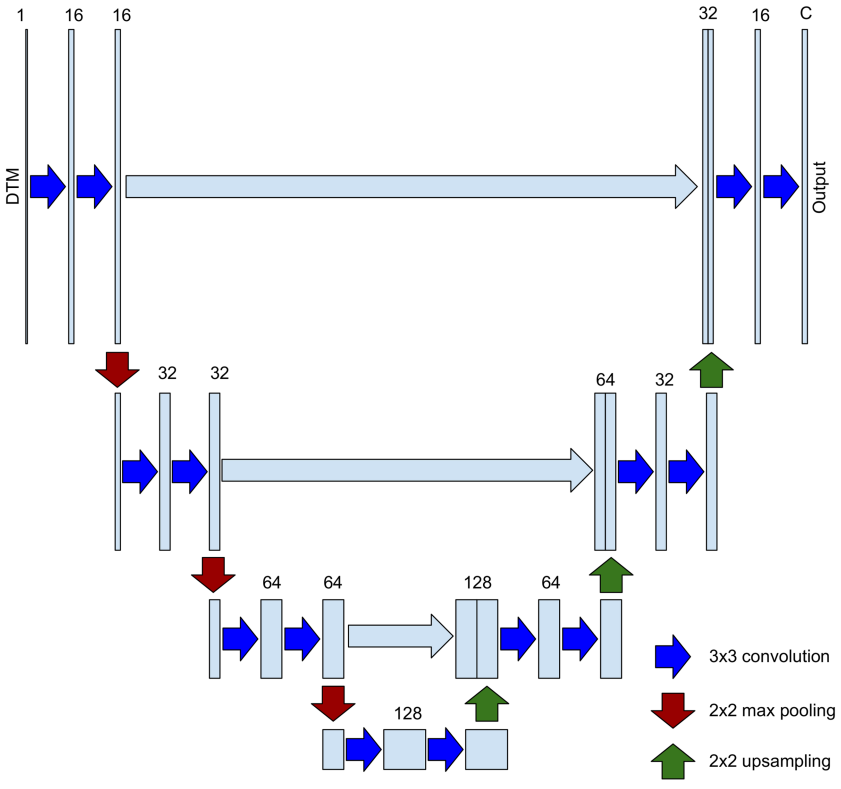

FIGURE 3 Architecture of U-Net as used in this study. The arrows indicate mathematical operations performed in the model in order to classify the input images into geomorphological maps. The light blue blocks represent the input and processed data, also called tensors. The number above the tensor indicates the number of channels in each tensor. The leftmost tensor is the DTM image, containing a single channel, while the last one is a tensor of the same spatial extent as the input but with *C* channels, where *C* refers to the number of output classes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Names of the three subsets, their classes and the selection criteria to derive these classes from the 2019 geomorphological map (see Figure 2 for details of the geomorphological units)

Subset	Name	Class number: name	Selection criteria
1	Aggregated landforms	1a: Push moraine	Geomorphological units B11 and D11
		1b: Other ridges	Landform B, excluding B11
		1c: Sloping areas	Landforms G and H
		1d: Complexes of small landforms	Landform L
		1e: Plains	Landform M
		1f: Non-valley-shaped depressions	Landform N
		1g: Valleys	Landform R
2	Dry valleys	2a: Dry valleys	Geomorphological unit R21
		2b: Background	All others
3	Drift sands	3a: Drift sands	Geomorphological units L54, N51, M54, B58 and units with addition 's'
		3b: Background	All others

typically up to 50 m in diameter. Nonetheless, the overall characteristics of the natural terrain were not modified to a large extent by these filters and the landforms of interest were still well visible. We then applied a Laplacian filter of 9 × 9 cells on the smoothed DTM to create a relief map. The DTM and the relief map served as input for U-Net. The input raster files were divided into tiles of 256 × 256 cells, with an overlap of 25% to improve predictions around the edges of the tiles.

The size of the raster cells determines the extent and level of detail of the geomorphological units that U-Net can recognize, because the number of cells in the tiles is fixed. Larger raster cells give a larger spatial extent on which larger geomorphological units might be more easily recognizable, but provide less detail. Conversely, smaller raster cells might work better to recognize smaller objects, but they provide a smaller spatial context as well. To test how raster resolution affected the results, we worked with three cell sizes representing different spatial extents (5, 10 and 25 m). All together this resulted in 27 model runs (three subsets of geomorphological units × three sets of weights × three cell sizes).

3.4 | Evaluation procedure

We performed a quantitative evaluation of all the maps created in this study in three steps (Figure 4):

1. The *manual maps* were compared among each other.
2. The *computed maps* were evaluated by comparison with the *2019 geomorphological map*.
3. The *manual* and *computed maps* were compared among each other and to the *consensus map*.

Raster files are easier to compare with each other and with the results of the semi-automated mapping. Therefore, the manual maps, that were mapped in vector format, were converted to raster format for evaluation. We used a resolution of 10 m, which roughly corresponds to a map scale of 1:20 000 (Tobler, 1987), which was the approximate scale that most geomorphologists used for their mapping.

We performed the evaluation of the manual maps on the geomorphological units in the map. We determined the agreement between the manual raster maps by stacking them and determining the most commonly mapped geomorphological unit for each raster cell. The frequency of this class determined the agreement between the manual maps, which ranges from 0 to 1 [Equation 1]:

$$\text{agreement} = \frac{(\text{frequency} - 1)}{(\#\text{maps} - 1)} \quad (1)$$

This procedure resulted in two maps: a *composite map* showing the most commonly mapped geomorphological unit (mode) and an *agreement map* showing the agreement between the different geomorphologists (Table 1). The agreement map was aggregated to a class agreement and overall agreement by spatially averaging the agreement map per class or over the entire map.

The maps created by the semi-automated mapping were evaluated by comparing them with the 2019 geomorphological map, which was also used to train the model. We selected two standard statistics to evaluate machine learning results. These are overall accuracy and Cohen's kappa. The overall accuracy (OA) describes the overall agreement between two maps, based on the cells that are classified the same in both maps, divided by the total number of cells [Equation 2]:

$$\text{OA} = \frac{\text{correctly classified cells}}{\text{total number of cells}} \quad (2)$$

The agreement determined with OA does not depend on the frequency of individual map components, which means that every raster cell has the same weight. The scores of the accuracies range from 0 to 1, where 1 indicates perfect agreement. Depending on the number of classes, a value of 0.125 (eight classes) or 0.5 (two classes) indicates a random agreement.

Cohen's kappa (κ) (Cohen, 1960) measures the agreement between two maps by correcting the OA with an expected random chance accuracy (p_e), meaning that it corrects for agreements that can occur by chance [Equation 3]. p_e is defined as the sum of the product of the number of times n class k is classified on map

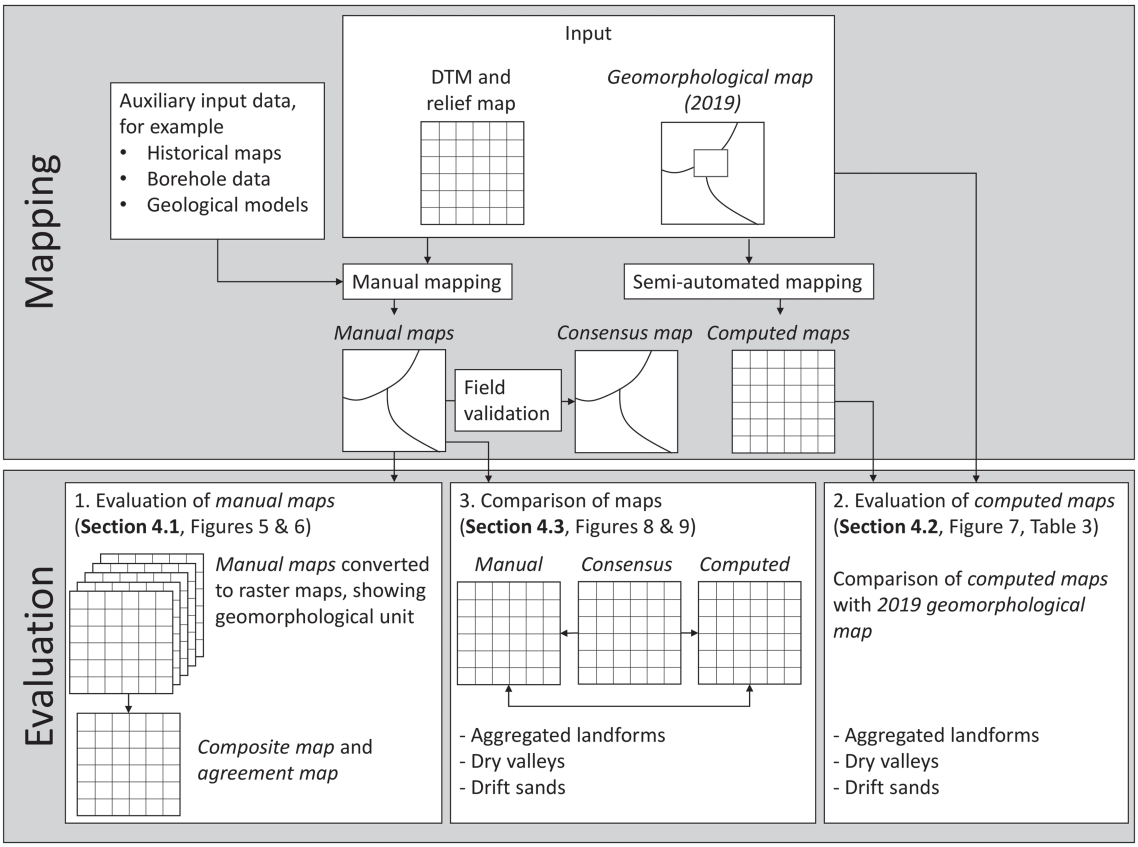

FIGURE 4 Workflow for making and evaluating the manually and semi-automatically created geomorphological maps. The tiles with curved lines represent maps in vector format. Tiles with grids represent maps in raster format. All analyses were performed with maps in raster format. Map names are detailed in Table 1

1 and map 2, divided by the square of the total number of observations N:

$$\kappa = \frac{OA - p_e}{1 - p_e}, \text{with } p_e = \frac{1}{N^2}\sum_k n_{k1} n_{k2} \quad (3)$$

The values of Cohen's kappa range from −1 to 1, where 1 indicates perfect agreement, 0 indicates an agreement that can be expected by random chance and values lower than zero indicate an agreement that is worse than random. Both statistics can handle multiple classes. Cohen's kappa also gives similar importance to less common classes as to more common classes. This is important, because local geomorphological units often cover a very small area relative to regional or background units, while they play a large role in landscape functions like biodiversity or water routing.

To test the effect of uncertainty in ground-truth data on the evaluation of the computed maps, we evaluated the computed maps using each manual map as a separate ground-truth map. This resulted in a set of evaluation statistics for each computed map, with one value for each of the manual maps. For this, the manual maps were converted to raster maps using the same classification as was used for making the subsets for training the semi-automated mapping (aggregated landforms, dry valleys and drift sands; Table 2).

Next to that, we also wanted to compare the quality of the manual maps and the computed maps compared to an independent map. For this, we selected the consensus map. The manual map on which this consensus map was based was left out of the further analysis. We calculated Cohen's kappa for each manual or computed map in comparison to the consensus map.

4 | RESULTS

4.1 | Evaluation of manual maps

The five manual maps that were developed in this study are presented in Figure 5. From visual qualitative analysis, each map sketches the same general geomorphological division of the study area, with the push moraine in the centre, flanked by coversands and drift sands to the east and sandrs and drift sands to the west. Dry valleys are oriented south-westwards off the push moraine on the western flanks, and are also present on the eastern slopes—albeit less prominent. There are also considerable differences, such as the length of the dry valleys and the extent of the drift sands in the western part of the map. Some geomorphologists made the drift sand in their maps less prominent than others, by mapping it as a legend unit addition instead of a geomorphological unit (cf. Figure 1), but these additions are not made visible in these maps. In practice this means that some geomorphologists made the drift sands more prominent than others. Another difference is the level of detail of the maps. The geomorphologists with the most experience with geomorphological mapping (M1, M4) drew the most individual polygons. The average number of vertices for each polygon also differs per map, with M1, M3 and M4 (working at Wageningen Environmental Research) being higher than M2 and M5 (working elsewhere).

The manually made maps were aggregated to the composite map visible in Figure 6a, which shows the most commonly mapped geomorphological unit (mode) on the manual maps. Figure 6b shows the agreement between the manual maps. The highest agreements can be

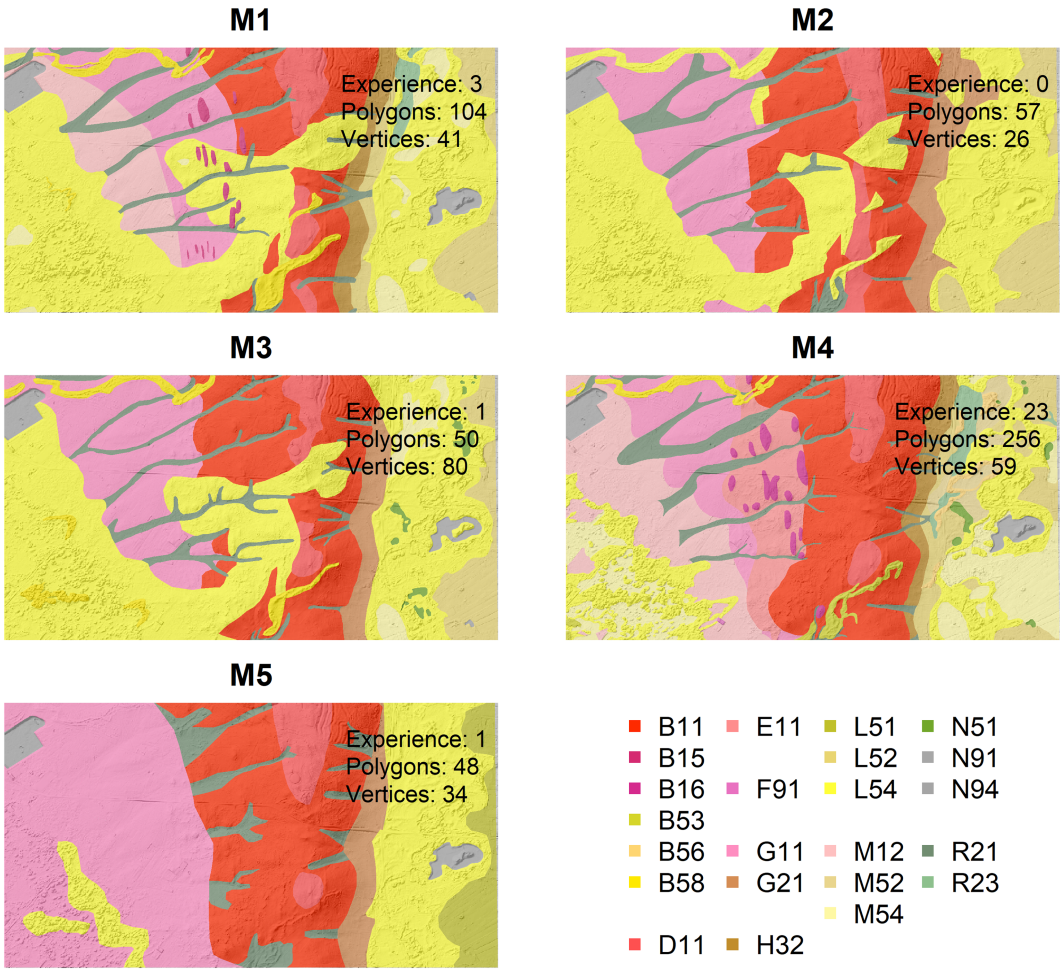

FIGURE 5 The main geomorphological units as mapped on the different manual maps (M1–M5). Map M1 is the 2019 geomorphological map. The text in the maps shows the years of experience with geomorphological mapping, the number of individual polygons in the map and the average amount of vertices per polygon. See Figure 2 for descriptions of the geomorphological units belonging to the different codes. The base map for each pane is a hillshade map of the DEM used in the study [Color figure can be viewed at wileyonlinelibrary.com]

found in the centre of larger geomorphological units, such as the push moraine (B11 and D11), sandrs (G11) and drift sand complexes (L54). Smaller distinct geomorphological units that are clearly distinguishable in the landscape are also mapped with higher agreement. Examples are dry valleys (R21) and quarries (N91). However, at the edges of these geomorphological units, both in width and length, the agreement decreases. This indicates that there is considerable variation on where the edges of the geomorphological units are drawn. For larger geomorphological units, this edge has a relatively smaller effect than for smaller geomorphological units. Geomorphological units with the lowest agreement are those which might be too small for the map scale and therefore only mapped by a few geomorphologists [e.g. ice-meltwater hills (B15)], or geomorphological units which can be classified in multiple classes due to limited information for correct classification [belted and non-belted coversands (L52 and L51)].

Figure 2b shows the consensus map which was created by incorporating findings from the field visit by all geomorphologists into the manual map M4. The drawn boundaries in this map most closely followed the boundaries as determined in the field, and contained the most detail as well. The field visit provided new insights into the genesis of several landforms. The most importance conclusion was that the dune area east of the push moraine was classified by all team members as a recent drift sand area (L54 or M54; Figure 5). Corings in the field, however, showed that these dunes contain well-developed dry Podzol soils, indicating that the soils must have been stable for several millennia. It was therefore jointly concluded that the dunes consist of Pleistocene/early Holocene coversands instead of recent drift sands. Another major result was that the extent of the drift sand area in the west turned out to be much smaller than on most manual maps. The region between the drift sand and the push moraine indeed has a locally small cover of drift sand, but the main geomorphological unit is the underlying flat sandr plain. Based on these findings in the field, it was concluded that the drift sand in that area is therefore better represented by the additional 's' on top of the main sandr plain, rather than classifying it as drift sand area.

4.2 | Evaluation of computed maps

For the analysis of the computed maps, we focus on the maps created using the weights calculated by the square root of the inverse distribution. This set of weights performed best overall for the different subsets. Quantitative evaluation results for all model runs are provided in Table S2.

Figure 7 shows a visual comparison of the computed maps and the ground-truth maps that were derived from the 2019

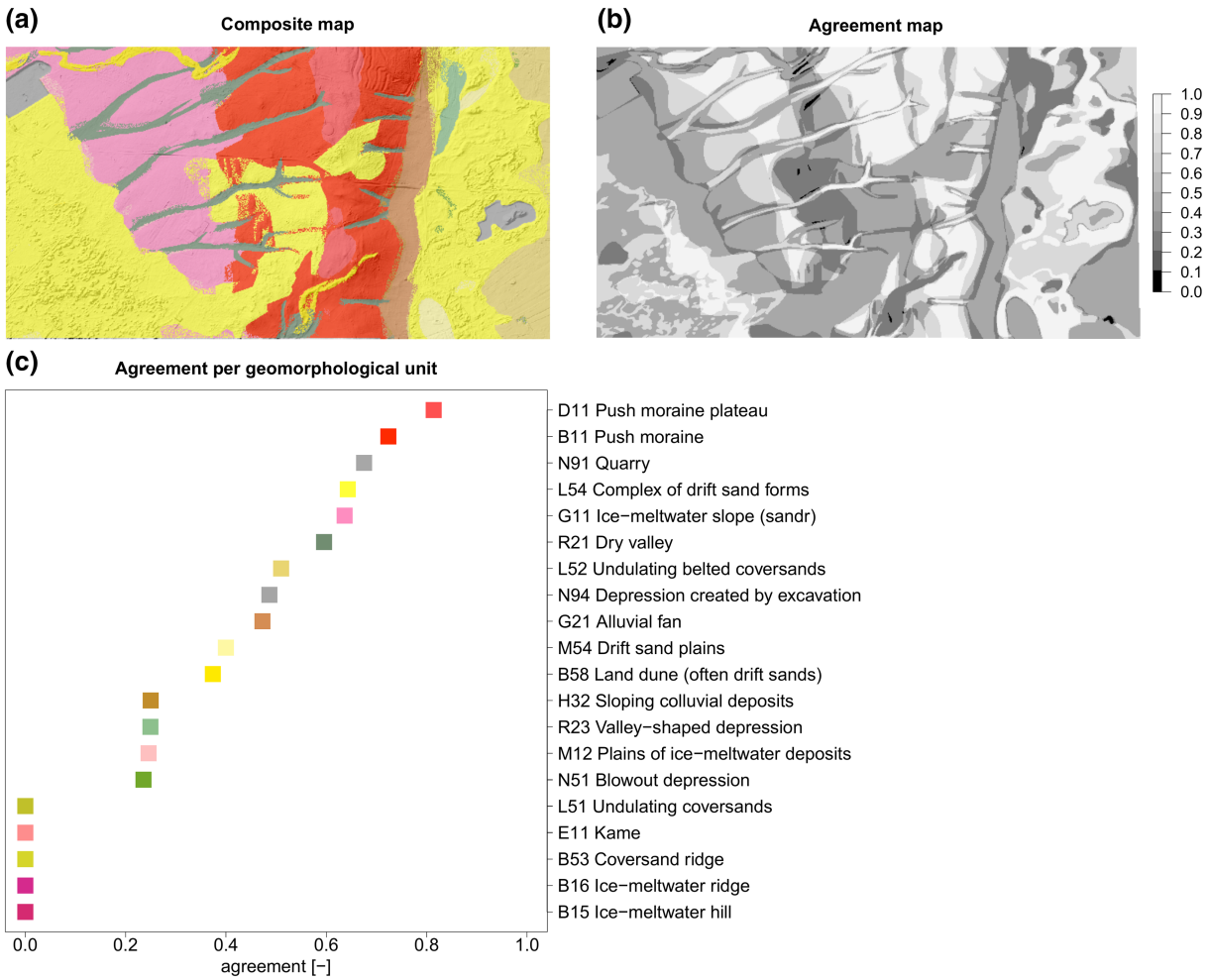

FIGURE 6 Composite map created from the manual maps showing the main geomorphological units (a), the agreement between the different maps (b) and the agreement of this map summarized per landform (c). See Figure 2 for more details on the legend items [Color figure can be viewed at wileyonlinelibrary.com]

geomorphological map for the aggregated landforms, dry valleys and drift sands. For the aggregated landforms, the general geographical position of each class is predicted correctly, with the dominant push moraine in the centre, flanked by the sandr and sandr plains including the complexes of drift sand dunes on both the eastern and western sections of the test area. However, the predicted extent of the push moraine (class 1) is much larger than observed and the sloping areas (class 3) and complex land forms (class 4) are predicted with different extents as well. Relatively large differences between the ground-truth and the predicted maps occur at a local scale as well: the small ridges (class 2) are predicted mainly at the wrong locations, while the dry valleys (class 7) seem not to be predicted at all.

For the dry valleys, which are topographically lower than their surroundings, mainly the steeper parts of the valleys on the push moraine are detected, while the valleys on the flatter sandr are often not recognized. The CNN performed reasonably well in detecting the typical elongated patterns of such valleys, but there are gaps in the predicted valleys. Also, the extent of the drift sand areas is predicted much smaller than how they were mapped on the ground-truth map. The predicted drift sands are mainly located in the areas with steeper topography and more pronounced dunes, such as the dune area to the east of the push moraine and in the southwest of the study area. Several single dunes on the flanks and the top of the push moraine were recognized as drift sand areas as well. In general, the semi-automated mapping performs reasonably well in detecting landforms, but is less precise in the correct delineation. A model that is trained on a single landform (dry valleys of drift sands) seemingly performs better than a model that is trained on a set of aggregated landforms, because the latter contains more variation between classes due to the higher number of classes.

Table 3 shows the evaluation statistics for the semi-automated mapping results by comparing each computed map with the 2019 geomorphological map. The overall accuracy shows the overall agreement between the test data and the model fit. The aggregated landforms show the lowest overall accuracy. The dry valleys show the highest overall accuracy, with an almost perfect agreement. The high values for the overall accuracy are related to the number and frequency of the classes in the data. For example, for the subset with dry valleys, there is only a small area covered by the valleys, while the background takes up a large surface area. This distribution results in a high overall accuracy, where there is no penalty for less represented classes. A comparison between overall accuracy from the training and test set can say something about model overfitting or underfitting. For the aggregated landforms and the dry valleys, the statistics are in the same order of magnitude (see Table S2). For the drift sands, however, the overall accuracy for the training set is much higher than for the test set, indicating overfitting of the model.

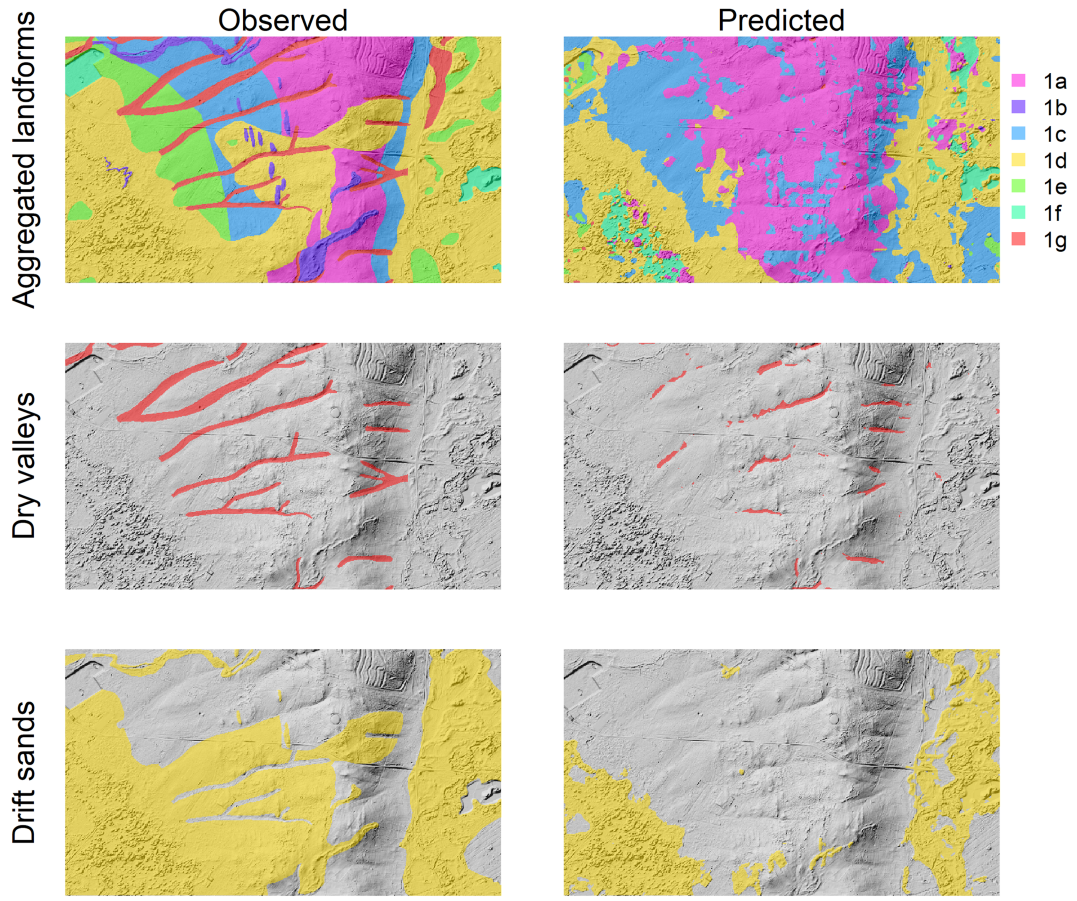

FIGURE 7 Overview of ground-truth (observed, left) and best-performing computed maps (predicted, right) for the different subsets of geomorphic units. Legend for the aggregated landforms: 1a, push moraine; 1b, other ridges; 1c, sloping areas; 1d, complex landforms; 1e, plains; 1f, non-valley-shaped depressions; 1g, valleys [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Evaluation statistics for the model performances in the areas, for the square root weights and for each evaluated raster cell size. Values lower than expected by random chance are printed in italics (see Table S2 for a detailed overview of all evaluation statistics)

	Aggregated landforms			Dry valleys			Drift sands		
Cell size [m]	5	10	25	5	10	25	5	10	25
Overall accuracy [-]	0.45	0.44	0.45	0.94	0.94	0.94	*0.49*	0.64	0.61
Cohen's kappa [-]	0.2	0.24	0.25	0.2	0.32	0.19	0.11	0.34	0.29

For Cohen's kappa, the values range from 0.11 to 0.34, indicating a slight to fair agreement (Table 3). The differences for each cell size are minor, but in general, the 5 m cell size shows the lowest agreements. For 10 and 25 m, the results are similar. Apparently, a cell size of 5 m results in an extent that is too small to recognize the typical topographical properties of the geomorphological units.

4.3 | Comparison of different mapping methods

Figure 8 shows a visual comparison of the manual agreement map with the best-performing computed map for the dry valleys and the drift sands. The location where the model predicted a dry valley matches with the higher agreement of the manual maps. The model recognized the large parts of the dry valleys to the west of the push moraine, as well as most of the smaller dry valleys on the eastern side of the push moraine. However, the smaller branches—including the origins of the dry valleys—were not recognized by the model. Overall, the CNN is rather conservative in recognizing the dry valleys and appears to not recognize the typical elongated shape on gently sloping areas.

The model mainly identified the areas with rougher topography as drift sand areas. These are the dune areas in the southwest of the study area and the dunes to the east of the push moraine. Also a few individual dunes on top of the push moraine were correctly recognized as drift sands. These patterns again match with the regions with highest agreement from the manual maps. However, the model did not recognize the more diffuse drift sand cover on top of the sandr as drift sands. This area was also mapped with a lower agreement on the agreement map, since they were also classified as plains of ice-meltwater deposits with a local cover of drift sands.

In Figure 7, the manual maps were compared with the 2019 geomorphological map as ground truth. This map was also used to train the model outside the test area. However, as we have shown in Figure 5, there can be considerable differences between manually mapped geomorphological maps. To illustrate the effect of uncertainty

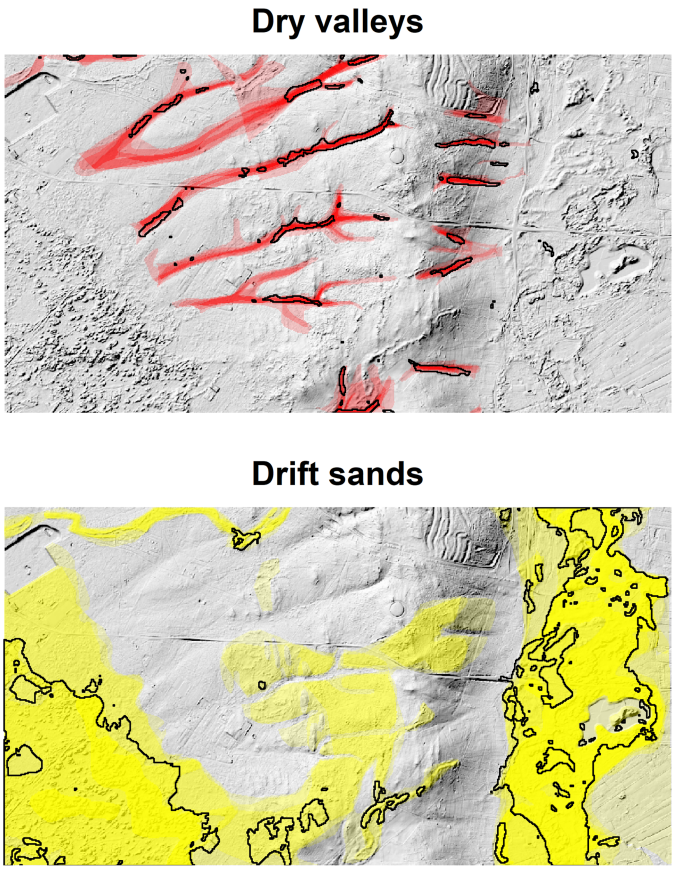

FIGURE 8 Comparison between manual mapping agreement of dry valleys (top) and drift sand areas (bottom) with the best-performing CNN prediction for both landforms delineated in black [Color figure can be viewed at wileyonlinelibrary.com]

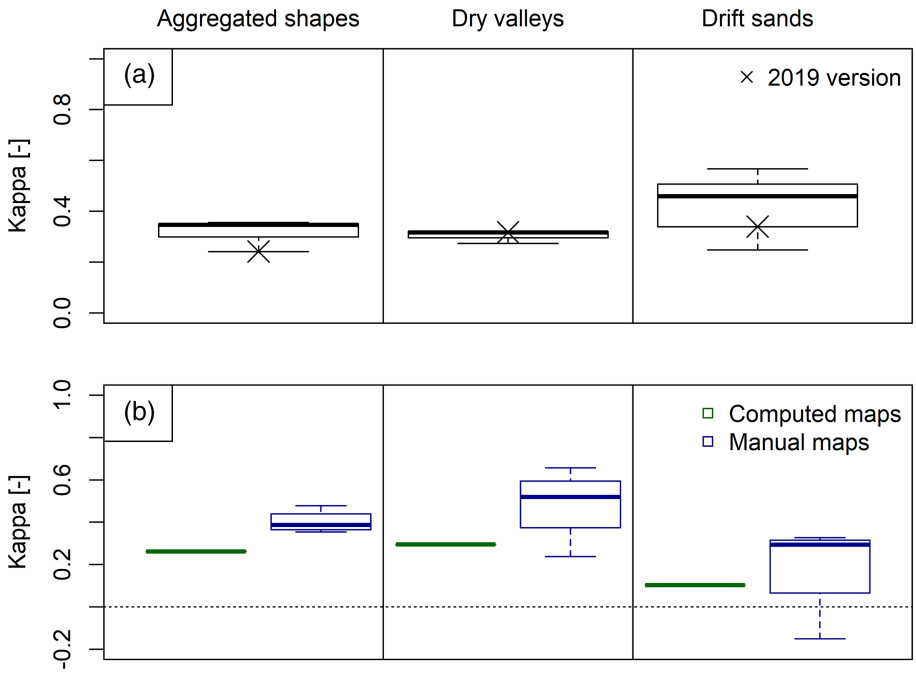

FIGURE 9 Boxplots showing the spread in Cohen's kappa (a) when the best computed map is compared to each separate manual map as ground truth and (b) when comparing the best computed map and all manual maps with the consensus map. The crosses in part A refer to the comparison where the 2019 geomorphological map was used as ground truth. This map was also used to train the data [Color figure can be viewed at wileyonlinelibrary.com]

in ground truth on the evaluation of semi-automated mapping, we used each manual map as ground truth for the computed maps and visualized the spread in Cohen's kappa in Figure 9a. There can be considerable spread in the calculated Cohen's kappa values, which range from 0.24 to 0.57. Remarkably, the comparison with the 2019 geomorphological map shows the highest agreement for the dry valleys, but for the other subsets most of the other manual maps show a higher agreement with the computed maps.

Figure 9b shows Cohen's kappa for all manual and best computed maps compared to the consensus map, which we consider to be the most representative ground-truth map due to the field validation. As mentioned, the manual map on which the consensus map was based was left out of this analysis. In general, the manual maps are more similar to the consensus map, with a few exceptions. In all cases, there is an overlap in the Cohen's kappa for the manual and computed maps. Of all subsets of geomorphological units, the dry valleys show

the highest similarity with the consensus map, closely followed by the aggregated landforms. The kappa scores range from 0.10 to 0.66, indicating a slight to fair agreement. There is one kappa value below zero for the drift sands, which indicates that any similarity between that map and the consensus map is lower than can be expected based on random chance. These maps are dissimilar rather than similar. The drift sands have the lowest similarity of each subset, probably because the extent of the drift sand area in the consensus map has changed considerably due to the field validation.

5 | DISCUSSION

5.1 | Quality of manual maps

When evaluating the quality of manual mapping, the results show that the general morphology of the area shows large similarities between each map, but there were also considerable differences between the spatial identification and classification of the geomorphological units (Figures 5 and 6) and delineation at smaller spatial scales.

Differences in geomorphological interpretation of the landscape can partly be attributed to the different backgrounds and working styles of the geomorphologists. In general, geomorphologists who maintain the geomorphological map as part of their daily job created maps with more polygons and more detail (Figure 5, M1, M3 and M4). Maps M2 and M5 were created by geomorphologists who were familiar with the geomorphological map, but did not have much experience with mapping. This reflects in the number of drawn polygons and the number of vertices per polygon. Also, different choices were made about which geomorphological units were prioritized. For example, in M5 the regional geomorphological units (sandr plains, drift sands and push moraine) were prioritized, while the other maps prioritized the local geomorphological units, such as the dry valleys, individual dunes and small ridges.

Another important issue coming out of the discussions was the use of scale levels. Formally, the application scale of the geomorphological map as part of the Dutch National Key Registry is defined as 1:50 000. This means that the map can be used for purposes at that scale, but is not to be used at more detailed scale levels without critical review. The scale level of 1:50 000 translates into a minimum size map unit of ~10 ha, which was favourable for printing purposes and readability of the printed maps (Steur et al., 1991). However, when mapping in GIS, often a more detailed scale can be used, because the currently available digital information facilitates mapping at a more detailed scale. In this project, although the mapping protocol was described in detail, no mapping scale levels were defined. This means that the level of detail in the manual maps exceeded the mapping scale of 1:50 000. Some of the geomorphologists used a self-imposed minimal scale, whereas others let this depend on the landform to be digitized. Overall, most maps were created at a scale of ~1:20 000. Also, the delineation of the specific landform played a role, where in some cases the landform can visually be sharply separated from bordering landforms, whereas in other cases the exact boundary is not relevant, since adjacent landforms can have diffuse boundaries and merge into one another at the application scale of the map. The differences in the maps due to different interpretations of the mapping scale indicate that a mapping scale should be defined explicitly in the mapping protocol. There can also be several mapping scales defined, depending on the geomorphological unit. Although the current data sources facilitate mapping at a very detailed scale (up to 1:1000), the applications of the final map often require less-detailed maps, which should be considered in defining the required level of detail. When different applications require different levels of detail from a geomorphological map, a hierarchical map and legend—that could be visualized and analysed at a wide range of scales—could offer a solution (De Jong et al., 2021).

The consistency, or inconsistency, of the manually created maps is also referred to as inter-rater reliability. Maps that are more similar have a higher inter-rater reliability, which results in higher evaluation statistics and higher agreement between different maps (McHugh, 2012). Ideally, the inter-rater reliability is as high as possible. This would indicate that the mapping technique of different mappers is very similar, which should result in a consistent map. In the agreement map in Figure 6b, this would be the case for parts of the push moraine, drift sands and the centres of the dry valleys. Lower inter-rater reliability would indicate that maps created by different geomorphologists are less consistent, for example the transition between push moraine and sandr in the west and between push moraine and the sand landscape in the west. A low inter-rater reliability could propagate into training datasets for semi-automated mapping exercises (Sadr, 2016). We discuss the inter-rater reliability in the manual maps in more detail below.

With mapping exercises like we performed in this study, areas and geomorphological units with a low mapping agreement could be pinpointed. In order to reduce differences in identification and delineation, scale and zooming rules should be implemented in the geomorphological mapping protocol. From the evaluation of the different maps and the validation in the field, it was also apparent that some legend units are more open to different interpretations than others. From the fieldwork, it became clear that general genesis was often agreed upon. Based on topography, there was high agreement on the sloping landforms. The interpretation on sandrs, alluvial fans or colluvial deposits showed less consistency. In addition, the dominance of one landform over another was sometimes interpreted differently. This was particularly the case in the more complex land units, where the discussion was often raised whether drift sand dunes were dominant over the underlying plains or undulating coversands, or whether the coversands should be regarded as more prominent than overlying small-scale land dunes. This is important, since a difference in interpretation might result in a seemingly different geomorphological coding, whereas the difference in meaning is much subtler. Additionally, differences in classifications were related back to boundary issues, especially where the border from one landform into the next was gradual both in the field and in the base data.

This has resulted in the similar general morphology of the area and the generally high agreement of the different maps (Figure 6). The field validation, however, showed that the agreement between the manual maps did not mean they were correct. Everyone mapped the dune area east of the push moraine as recent drift sands based on the high relief and the 2019 geomorphological map, while field work showed that these dunes should be classified as coversand dunes. In the legend of the geomorphological map, high relief is defined by slopes steeper than 1° (Maas et al., 2021). This shows that conventions about typical relief of geomorphological units are not always

valid, and field validation and continuous discussion should be an essential part of the protocol for large mapping campaigns to avoid mistakes due to wrong assumptions and mapping habits (Knight et al., 2011). These points were also included in the mapping protocol for the geomorphological map of The Netherlands (Van der Meij & Maas, 2020).

5.2 | Geomorphological mapping using CNNs

5.2.1 | Geomorphological mapping with CNNs in The Netherlands

Our attempt to map geomorphological units in the Dutch push-moraine district using a CNN was partly successful. The model results showed that, based on the surrounding training areas, a reasonable agreement was achieved when mapping the aggregated landforms, as well as single geomorphological units (dry valleys or drift sands; Table 3). Still, the models were mainly successful in identification of parts of the targeted geomorphological units instead of the entire unit. For the dry valleys, the models identified parts of the valleys, but did not predict a continuous valley from the push moraine into the sandr areas. The morphology of the dry valleys on the push moraine is much steeper and deeper incised than the valleys on the sandr, which leads to a relative conservative interpretation by the CNN. Additionally, the CNN is not able to recognize the entire typical elongated shape of the legend item, and fails to connect upslope and downslope portions of the same dry valley. In the legend of the geomorphological map, these parts of the valleys are distinguished by the relief code, but in the models no distinction was made. These ambiguous properties of the geomorphological unit might have contributed to the moderate results of the model. The same goes for the drift sand areas. The models predicted the drift sands mainly in areas with more pronounced dune morphology. These areas are identified as a complex of land dunes in the legend of the geomorphological map. Areas with less pronounced drift sand morphology were not recognized by the model, although these areas were present in the training data. Based on field observations and corings, we decided that these areas should not be classified as drift sands as the main geomorphological unit in the consensus map, but as ice-meltwater deposits (sandr) with a cover of drift sands (Figure 2b). The morphology was thus leading in the delineation of drift sands by the CNNs as well as by the geomorphologists. This again emphasized the need for field observations (by multiple geomorphologists) as part of updating geomorphological maps, both by manual mapping and semi-automated mapping (Knight et al., 2011).

Studies that aimed to map specific landforms using CNNs achieve overall better results than our study. For example, a study mapping ice-wedge polygons in Alaska correctly predicted 91% of the polygons, but also noted that their algorithm was unable to map polygons with low relief (Abolt & Young, 2020). Palafox et al. (2017) mapped geological landforms on Mars with accuracy around 94%. Verschoof-Van der Vaart and Lambers (2019) achieved F1 scores of 0.68 and 0.79 for detecting archaeological landforms in The Netherlands. For comparison, the F1 scores from our study ranged between 0.08 and 0.71 (see Table S2). Studies using other semi-automatic mapping techniques for mapping the Dutch geomorphology yield similar results as our CNN. For example, Seijmonsbergen et al. (2011) mapped low-relief landforms in the northern Netherlands using object-based image analysis. Their results match relatively well with the original geomorphological map of that area, with kappa values of around 0.57. Despite the higher kappa values, their models also did not accurately map specific shapes of geomorphological units, such as elongated valleys or dune ridges. This comparison indicates that our model scores lower than other studies using similar methods. We partly contribute this to the relatively large sizes of the geomorphological units that we want to predict and to the anthropogenic structures (roads, ditches) that are still in the DTM, despite the filtering. Such structures are absent in the pristine landscapes of Alaska and Mars. Nonetheless, we see room for improvement when mapping geomorphological units using CNNs.

5.2.2 | Recommendations for geomorphological mapping with CNNs

There are several steps that could be taken to improve semi-automatic geomorphological mapping using CNNs. These are: (1) using additional data sources as training data; (2) post-processing of the results; and (3) exploring a combination with other methods for semi-automatic geomorphological mapping.

In this first attempt at geomorphological mapping in The Netherlands using CNNs, we only used a DTM and a relief map as training data, and relied on the self-learning capabilities of CNNs to recognize the geomorphological units on the training data. To improve the predictions, there are data sources that show other terrain or landscape properties, which can be used as additional training data. These training data can, for example, be derived from the DTM. Such LSPs or geomorphometric variables (Hengl & Reuter, 2008) can, for example, contain information about the local relief (roughness) or the hydrological situation (topographic wetness index or flow accumulation). These LSPs provide information about the topography that is not readily available in DTMs or relief maps. Also, this information is difficult or impossible to derive by CNNs, because they require calculations over a larger spatial extent than that covered by the moving window calculations done in the CNNs. In landscapes with limited relief, such as coastal regions like the western Netherlands, the relief of anthropogenic structures—such as roads, ditches and buildings—is often much more pronounced than the natural relief. A simple filter such as we applied would not suffice to remove these anthropogenic objects from the DEMs; extensive filtering is required (e.g. Meyer-Heß, 2020; Van der Meulen et al., 2020). In built-up areas, the natural relief might even be completely overprinted and not be retrievable at all by DEM filtering, and data sources other than DTMs and LSPs are required.

In these cases, and in cases where different geomorphological units have similar morphology, a distinction can be made based on genesis. Examples of these geomorphological units are Holocene and Pleistocene land dunes as shown in this study, or fluvial and marine meander belts and channels which are omnipresent in many near-coastal regions in the world. Information on genesis can be included via remotely sensed spectral maps that contain information on soil moisture or soil texture (Castaldi et al., 2016; Chan et al., 2016). Such maps perform well in digital soil mapping (Maynard & Levi, 2017). Other useful maps that can be used to separate geomorphological

units on their genesis are continuous (geological) probability maps of the subsurface's texture or hydraulic conductivity, such as GeoTOP and REGIS in The Netherlands (Stafleu & Dubelaar, 2016). These maps are often based on coring data and directly represent the soil properties, instead of via a correlation like with spectral maps. In addition, using coring data themselves may also improve CNN performance. These geological maps and corings were also used during the manual mapping.

Post-processing can also greatly improve the quality of the predictions from CNNs and other semi-automatic mapping techniques (Salvi et al., 2021). Operations such as smoothing and cleaning can remove noise from the predictions and create more realistic shapes for the predictions (Sarzana et al., 2020). This could, for example, benefit the mapping of elongated, continuous shapes of the dry valleys or ridges. Also using more advanced techniques, such as conditional random fields, could greatly improve the results of CNNs by providing sharper boundaries of the mapped objects (Arnab et al., 2018).

Finally, it will also be useful to explore a combination of different techniques for semi-automatic mapping. The results from our study show that predictions for single geomorphological units with CNNs are more realistic than spatially explicit maps, and that each geomorphological unit scores best with different weights and cell sizes (Table 3). Other techniques, such as object-based image analysis or classification by random forests, have been successful in mapping specific landforms or creating spatially explicit maps (Anders et al., 2011; Dou et al., 2015; Giaccone et al., 2021; Pedersen, 2016; Seijmonsbergen et al., 2011). This suggests that when different geomorphological units are mapped using the most suitable algorithm or parameter set, the results can be combined or stacked into a single, accurate geomorphological map.

5.3 | Effect of uncertainty in ground-truth data on model performance

In machine learning, the ground-truth data (i.e. data used to verify the model results) cannot be assumed to be without error (Bowler et al., 2020; Zhang et al., 2020). The ground-truth data are annotated or labelled by humans. This uncertainty in ground-truth data is often not considered in machine learning. In studies where they do address this issue, they often find (unacceptably) high variation in classification by different individuals (Lyman & VanPool, 2009; Sadr, 2016), although errors in the test data might also help to identify more robust models (Northcutt et al., 2021).

In our study we also observe a high inter-rater variability in the manual maps (Figures 5 and 8). When we used the different manual maps as test data for the CNN results, the evaluation statistics could vary widely (Figure 9a). In our case, the model performance ranged from poor to moderate, depending on the subset and ground truth. For the aggregated landforms and drift sands, the 2019 geomorphological map (the training dataset) does not show the highest agreement with the computed maps in the test area. In fact, most of the other manual maps have a higher agreement with the computed maps. Figure 9a shows that uncertainty in ground truth can have a substantial effect on model evaluation and might affect the selection of the best-performing model. This uncertainty not only plays a role in the evaluation of the models. During the training phase, the ground-truth maps also contain a certain degree of uncertainty that will affect the performance of the model. When the training data are annotated by multiple people, there can be systematic differences in the ground-truth data that can lead to poorer performance of the models (e.g. Sadr, 2016). This shows the need for proper evaluation of the ground truth-data before training and evaluating the models, because the ground-truth data cannot be assumed to be without error.

The borders between different geomorphological units are often diffuse, which creates uncertainty when deciding where to draw the border. This partly depends on the intended scale level of the map and the scale level at which the mapping was performed. Also, this does not mean that the map is incorrect, since the aim of geomorphological maps is often to provide an interpretation of the origin and genesis of the landscape in consideration. It is important that the user of the map is aware of the aim of the map and the related uncertainty in the location of its boundaries. Moreover, the legend with which the landforms are classified is also subject to interpretation, where the geomorphological units and their respective scale and borders are predefined: they can be interpreted differently by geomorphologists with different backgrounds, levels of experience and working styles. These sources of uncertainty may propagate into the models and the final results and might be one of the causes of the poor to moderate model performance.

5.4 | Implications for geomorphological mapping

This study showed that uncertainty is inherent in geomorphological mapping, due to different interpretations of the geomorphological legend, subjective choices during the mapping and different levels of experience of the geomorphologist.

For semi-automated geomorphological mapping, CNNs require a considerable amount of ground-truth data for training the models. This means that manual maps are a necessity at this stage. The geomorphological mapping of the Utrechtse Heuvelrug as training data, data preparation for the model and training of the CNNs were time-intensive tasks. Thus far, the use of CNNs has not saved us any time in our mapping efforts. Nonetheless, the models that were trained on specific landforms, like dry valleys and dune areas, can be applied in other regions where these landforms occur. In The Netherlands, and other regions of the world, there are ample landscapes where dry valleys and drift sands occur, and where geomorphological maps are not available or require updating. In these areas, the CNNs developed in this study can assist in the mapping efforts. The use of crowdsourcing has proven to be effective to generate training data or improve the predictions of CNNs (e.g. Herfort et al., 2019; Lambers et al., 2019). However, we do not foresee that such an approach might be useful for geomorphological mapping, because it requires geomorphological training to read the landscape and correctly identify geomorphological units. Especially for regional geomorphological units with diffuse boundaries, such as sandr plains and drift sand areas, the delineation and classification might not be straightforward, as we have shown in this study with experienced geomorphologists.

At this stage, the incomplete delineation of the individual geomorphological units and the high uncertainty in the ground-truth data indicate that CNNs may only be used for detection of large landforms

or distinct geomorphological units instead of complete delineation and mapping of these units. The human geomorphologist remains necessary to properly delineate and classify the geomorphological units that are predicted by the models. At this stage, semi-automated mapping with CNNs can be seen as an additional tool in the mapping process, as the computed maps can identify areas that are ambiguously defined or may spark the discussion of classification definitions, scale issues and landform dominance. This is a more common find in studies that try to detect landforms with deep learning (e.g. Verschoof-Van der Vaart & Lambers, 2021). Before CNNs can (partly) take over the work of manual mappers in creating geomorphological maps and become part of the workflow for geomorphological mapping, more development and testing of the models, parameters and input is required. This will be the focus of future research.

6 | CONCLUSIONS AND RECOMMENDATIONS

Geomorphological mapping is a time-intensive task that is often difficult to reproduce due to subjective choices during the mapping. Semi-automated mapping techniques might deliver more robust and reproducible results. The objective of this study was threefold: (1) we wanted to quantify the uncertainty of manual geomorphological mapping; (2) we wanted to explore the use of CNNs for semi-automated geomorphological mapping; and (3) we wanted to test how the uncertainty in manual maps affected the quality of computed maps.

The manual maps created in this study show high similarities on a regional level, but can differ substantially at a local level, due to differences in experience with mapping, choices in delineation, differences in working styles and different interpretations and classifications of geomorphological units. Agreement between manual maps does not necessarily mean that the maps are correct (i.e. that they show the actual geomorphology of the landscape). Coordination of mapping efforts and field validation remain necessary to create accurate and precise geomorphological maps.

The CNNs in this study were mainly successful in identifying the general location of landforms and geomorphological units, but failed at correct delineation. The predictions might be improved by adding data layers that contain information about substrate and genesis, by (advanced) post-processing of the results or by combining different semi-automatic mapping techniques.

The quality of the computed maps varies substantially when using different manual map for evaluation. This shows that ground-truth data cannot be assumed to be without error, and that coordinated mapping is required when developing accurate and precise training and test data for semi-automated mapping techniques.

Our attempt at geomorphological mapping using CNNs as a semi-automatic mapping technique shows that, at this stage, CNNs can be a useful aid in geomorphological mapping, by identifying landforms, but also in discussing and further defining the classification of landforms. The human geomorphologist, however, remains necessary to correct the delineation and classification of the model results. More development on the models, data processing and parameters is required before CNNs can become a standalone technique for geomorphological mapping.

ACKNOWLEDGEMENTS

The authors thank Kim Cohen for his constructive comments during the manual mapping campaign and the quality assessment of the geomorphological map of The Netherlands. We thank Daniel van der Maas of Ellipsis Drive for developing the first versions of U-Net for geomorphological mapping in The Netherlands. We thank Joop Okx for coming up with the idea of using machine learning for semi-automated geomorphological mapping. Lastly, we thank two anonymous reviewers for their constructive feedback on the manuscript.

This publication presents the findings of research for the Statutory Research Tasks Unit for Nature and the Environment (WOT Natuur & Milieu) and is partly funded by the Dutch Ministry of Agriculture, Nature and Food Quality (LNV).

DATA AVAILABILITY STATEMENT

Our version of U-Net and the trained models are available on request to the first author. The findings of this study will be incorporated in the geomorphological map of The Netherlands, which is freely available via https://broloket.nl/.

ORCID

W. Marijn van der Meij https://orcid.org/0000-0001-8724-5120
Erik W. Meijles https://orcid.org/0000-0001-8738-1907
Diego Marcos https://orcid.org/0000-0001-5607-4445
Tom T. L. Harkema https://orcid.org/0000-0003-4922-5078
Jasper H. J. Candel https://orcid.org/0000-0001-8365-8673
Gilbert J. Maas https://orcid.org/0000-0002-7622-9499

REFERENCES

Abolt, C.J. & Young, M.H. (2020) High-resolution mapping of spatial heterogeneity in ice wedge polygon geomorphology near Prudhoe Bay, Alaska. *Scientific Data*, 7(1), 87. https://doi.org/10.1038/s41597-020-0423-9

Albawi, S., Mohammed, T.A. & Al-Azawi, S. (2017). Understanding of a convolutional neural network. In: *Proceedings of the 2017 International Conference on Engineering and Technology (ICET)*. Antalya, Turkey, pp. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Anders, N.S., Seijmonsbergen, A.C. & Bouten, W. (2011) Segmentation optimization and stratified object-based analysis for semi-automated geomorphological mapping. *Remote Sensing of Environment*, 115(12), 2976–2985. https://doi.org/10.1016/j.rse.2011.05.007

Arnab, A., Zheng, S., Jayasumana, S., Romera-Paredes, B., Larsson, M., Kirillov, A. *et al.* (2018) Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Processing Magazine*, 35(1), 37–52. https://doi.org/10.1109/MSP.2017.2762355

Aspinall, R.J. & Pearson, D.M. (1995) Describing and managing uncertainty of categorical maps in GIS. *Innovations in GIS*, 2, 71–83.

Bhuiyan, M.A.E., Witharana, C., Liljedahl, A.K., Jones, B.M., Daanen, R., Epstein, H.E. *et al.* (2020) Understanding the effects of optimal combination of spectral bands on deep learning model predictions: A case study based on permafrost tundra landform mapping using high resolution multispectral satellite imagery. *Journal of Imaging*, 6(9), 97. https://doi.org/10.3390/jimaging6090097

Bishop, M.P., James, L.A., Shroder, J.F. & Walsh, S.J. (2012) Geospatial technologies and digital geomorphological mapping: Concepts, issues and research. In: *Proceedings of the 41st Annual Binghamton Geomorphology Symposium*. Vol. 137. Columbia, SC, pp. 5–26. https://doi.org/10.1016/j.geomorph.2011.06.027

Bocco, G., Mendoza, M. & Velázquez, A. (2001) Remote sensing and GIS-based regional geomorphological mapping—a tool for land use planning in developing countries. *Geomorphology*, 39(3–4), 211–219. https://doi.org/10.1016/S0169-555X(01)00027-7

Bowler, E., Fretwell, P.T., French, G. & Mackiewicz, M. (2020) Using deep learning to count albatrosses from space: Assessing results in light of ground truth uncertainty. *Remote Sensing*, 12, 2026. https://doi.org/10.3390/rs12122026

BRO. (2021) About the key registry. Available at https://basisregistratieondergrond.nl/english/about-key-registry/ [accessed 17 June 2021].

Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S. & Casa, R. (2016) Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sensing of Environment*, 179, 54–65. https://doi.org/10.1016/j.rse.2016.03.025

Chan, S.K., Bindlish, R., O'Neill, P.E., Njoku, E., Jackson, T., Colliander, A. *et al*. (2016) Assessment of the SMAP passive soil moisture product. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4994–5007. https://doi.org/10.1109/TGRS.2016.2561938

Chelli, A., Bordoni, M., Cappadonia, C., Pepe, G., Rotigliano, E. & Smith, M. (2021) Geomorphological tools for mapping natural hazards. *Journal of Maps*, 17(3), 1–4. https://doi.org/10.1080/17445647.2021.1920794

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Corsini, A., Cervi, F. & Ronchetti, F. (2009) Weight of evidence and artificial neural networks for potential groundwater spring mapping: An application to the Mt. Modino area (Northern Apennines, Italy). *Geomorphology*, 111(1–2), 79–87. https://doi.org/10.1016/j.geomorph.2008.03.015

De Jong, M.G.G., Sterk, H.P., Shinneman, S. & Seijmonsbergen, A.C. (2021) Hierarchical geomorphological mapping in mountainous areas. *Journal of Maps*, 17(2), 214–224. https://doi.org/10.1080/17445647.2021.1897047

Deng, L., Hinton, G. & Kingsbury, B. (2013) New types of deep neural network learning for speech recognition and related applications: an overview. In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, pp. 8599–8603. https://doi.org/10.1109/ICASSP.2013.6639344

Dou, J., Chang, K.-T., Chen, S., Yunus, A.P., Liu, J.-K., Xia, H. & Zhu, Z. (2015) Automatic case-based reasoning approach for landslide detection: Integration of object-oriented image analysis and a genetic algorithm. *Remote Sensing*, 7(4), 4318–4342. https://doi.org/10.3390/rs70404318

Douglas, I. (2020) Urban geomorphology. In: Douglas, I., Anderson, P.M.L., Goode, D., Houck, M.C., Maddox, D., Nagendra, H. & Yok, T.P. (Eds.) *The Routledge Handbook of Urban Ecology*. Abingdon: Routledge.

Drăguţ, L. & Blaschke, T. (2006) Automated classification of landform elements using object-based image analysis. *Geomorphology*, 81(3–4), 330–344. https://doi.org/10.1016/j.geomorph.2006.04.013

Dramis, F., Guida, D. & Cestari, A. (2011) Nature and aims of geomorphological mapping. In: Smith, M.J., Paron, P. & Griffiths, J.S. (Eds.) *Developments in Earth Surface Processes: Geomorphological Mapping*. Amsterdam: Elsevier, pp. 39–73. https://doi.org/10.1016/B978-0-444-53446-0.00003-3

Du, L., You, X., Li, K., Meng, L., Cheng, G., Xiong, L. & Wang, G. (2019) Multi-modal deep learning for landform recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 63–75. https://doi.org/10.1016/j.isprsjprs.2019.09.018

European Commission. (2021) INSPIRE. Available at https://inspire.ec.europa.eu/ [accessed 7 January 2021].

Feizizadeh, B., Kazemi Garajeh, M., Blaschke, T. & Lakes, T. (2021) An object based image analysis applied for volcanic and glacial landforms mapping in Sahand Mountain, Iran. *Catena*, 198, 105073. https://doi.org/10.1016/j.catena.2020.105073

Giaccone, E., Oriani, F., Tonini, M., Lambiel, C. & Mariéthoz, G. (2021) Using data-driven algorithms for semi-automated geomorphological mapping. *Stochastic Environmental Research and Risk Assessment*. https://doi.org/10.1007/s00477-021-02062-5

Hengl, T. & Reuter, H.I. (2008) *Geomorphometry: Concepts, Software, Applications, Developments in Soil Science*. Amsterdam: Elsevier.

Herfort, B., Li, H., Fendrich, S., Lautenbach, S. & Zipf, A. (2019) Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. *Remote Sensing*, 11(15), 1799. https://doi.org/10.3390/rs11151799

Jongmans, A.G., Van den Berg, M.W., Sonneveld, M.P.W., Peek, G.J.W.C. & van Saparoea, R.M. (2015) *Landschappen van Nederland: Geologie, bodem en landgebruik*. Wageningen: Academic Publishers.

Kassouk, Z., Thouret, J.-C., Gupta, A., Solikhin, A. & Liew, S.C. (2014) Object-oriented classification of a high-spatial resolution SPOT5 image for mapping geology and landforms of active volcanoes: Semeru case study, Indonesia. *Geomorphology*, 221, 18–33. https://doi.org/10.1016/j.geomorph.2014.04.022

Knight, J., Mitchell, W.A. & Rose, J. (2011) Geomorphological field mapping. In: Smith, M.J., Paron, P. & Griffiths, J.S. (Eds.) *Developments in Earth Surface Processes: Geological Mapping*. Amsterdam: Elsevier, pp. 151–187. https://doi.org/10.1016/B978-0-444-53446-0.00006-9

Koomen, A.J.M. & Maas, G.J. (2004) *Geomorfologische kaart Nederland (GKN); achtergronddocument bij het landsdekkende digitale bestand*, Vol. No. 1039. Wageningen: Alterra.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2017) ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. https://doi.org/10.1145/3065386

Lambers, K., Verschoof-Van der Vaart, W.B. & Bourgeois, Q.P.J. (2019) Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection. *Remote Sensing*, 11(7), 794. https://doi.org/10.3390/rs11070794

Lark, R.M., Thorpe, S., Kessler, H. & Mathers, S.J. (2014) Interpretative modelling of a geological cross section from boreholes: Sources of uncertainty and their quantification. *Solid Earth*, 5(2), 1189–1203. https://doi.org/10.5194/se-5-1189-2014

Li, S., Xiong, L., Tang, G. & Strobl, J. (2020) Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery. *Geomorphology*, 354, 107045. https://doi.org/10.1016/j.geomorph.2020.107045

Lyman, R.L. & VanPool, T.L. (2009) Metric data in archaeology: A study of intra-analyst and inter-analyst variation. *American Antiquity*, 74(3), 485–504. https://doi.org/10.1017/S0002731600048721

Ma, L., Fu, T., Blaschke, T., Li, M., Tiede, D., Zhou, Z. *et al*. (2017) Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. *ISPRS International Journal of Geo-Information*, 6(2), 51. https://doi.org/10.3390/ijgi6020051

Maarleveld, G.C. & Van der Schans, R.P.H.P. (1961) *De dekzandmorfologie van de Gelderse Vallei*. Utrecht: KNAG.

Maas, G.J., Van der Meij, W.M., Van Delft, S.P.J. & Heidema, N. (2021) Toelichting bij de legenda Geomorfologische kaart van Nederland 1: 50 000. Available at https://legendageomorfologie.wur.nl/ [accessed 17 May 2021].

Maynard, J.J. & Levi, M.R. (2017) Hyper-temporal remote sensing for digital soil mapping: Characterizing soil–vegetation response to climatic variability. *Geoderma*, 285, 94–109. https://doi.org/10.1016/j.geoderma.2016.09.024

McHugh, M.L. (2012) Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. https://doi.org/10.11613/BM.2012.031

Meyer-Heß, M.F. (2020) Identification of archaeologically relevant areas using open geodata. *KN Journal of Cartography and Geographic Information*, 70(3), 107–125. https://doi.org/10.1007/s42489-020-00049-w

Northcutt, C.G., Athalye, A. & Mueller, J. (2021) Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv: 2103.14749.

Palafox, L.F., Hamilton, C.W., Scheidt, S.P. & Alvarez, A.M. (2017) Automated detection of geological landforms on Mars using convolutional

neural networks. *Computers & Geosciences*, 101, 48–56. https://doi.org/10.1016/j.cageo.2016.12.015

Paron, P. & Claessens, L. (2011) Makers and users of geomorphological maps. In: Smith, M.J., Paron, P. & Griffiths, J.S. (Eds.) *Developments in Earth Surface Processes: Geomorphological Mapping*. Amsterdam: Elsevier, pp. 75–106. https://doi.org/10.1016/B978-0-444-53446-0.00004-5

Pedersen, G.B.M. (2016) Semi-automatic classification of glaciovolcanic landforms: An object-based mapping approach based on geomorphometry. *Journal of Volcanology and Geothermal Research*, 311, 29–40. https://doi.org/10.1016/j.jvolgeores.2015.12.015

Pierik, H.J. & Cohen, K.M. (2020) The use of geological, geomorphological and soil mapping products in palaeolandscape reconstructions for the Netherlands. *Netherlands Journal of Geosciences*, 99, 1–20. https://doi.org/10.1017/njg.2020.8

Pierik, H.J., Van Lanen, R.J., Gouw-Bouman, M.T., Groenewoudt, B.J., Wallinga, J. & Hoek, W.Z. (2018) Controls on late-Holocene drift-sand dynamics: The dominant role of human pressure in the Netherlands. *The Holocene*, 28(9), 1361–1381. https://doi.org/10.1177/0959683618777052

Randle, C.H., Bond, C.E., Lark, R.M. & Monaghan, A.A. (2018) Can uncertainty in geological cross-section interpretations be quantified and predicted? *Geosphere*, 14(3), 1087–1100. https://doi.org/10.1130/GES01510.1

Roccati, A., Mandarino, A., Perasso, L., Robbiano, A., Luino, F. & Faccini, F. (2020) Large-scale geomorphology of the Entella River floodplain (Italy) for coastal urban areas management. *Journal of Maps*, 17(4), 98–112. https://doi.org/10.1080/17445647.2020.1738281

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Strasbourg, France. Springer, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sadr, K. (2016) The impact of coder reliability on reconstructing archaeological settlement patterns from satellite imagery: A case study from South Africa. *Archaeological Prospection*, 23(1), 45–54. https://doi.org/10.1002/arp.1515

Salvi, M., Acharya, U.R., Molinari, F. & Meiburger, K.M. (2021) The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128, 104129. https://doi.org/10.1016/j.compbiomed.2020.104129

Sarzana, T., Maltese, A., Capolupo, A. & Tarantino, E. (2020) Post-processing of pixel and object-based land cover classifications of very high spatial resolution images. In: Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D. *et al.* (Eds.) *Computational Science and Its Applications – ICCSA 2020*, *Lecture Notes in Computer Science*. Cham: Springer International, pp. 797–812. https://doi.org/10.1007/978-3-030-58811-3_57

Seijmonsbergen, A.C., Hengl, T. & Anders, N.S. (2011) Semi-automated identification and extraction of geomorphological features using digital elevation data. In: Smith, M.J., Paron, P. & Griffiths, J.S. (Eds.) *Developments in Earth Surface Processes: Geomorphological Mapping*. Amsterdam: Elsevier, pp. 297–335. https://doi.org/10.1016/B978-0-444-53446-0.00010-0

Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M. & Bajat, B. (2020) Random forest spatial interpolation. *Remote Sensing*, 12, 1687. https://doi.org/10.3390/rs12101687

Shumack, S., Hesse, P. & Farebrother, W. (2020) Deep learning for dune pattern mapping with the AW3D30 global surface model. *Earth Surface Processes and Landforms*, 45(11), 2417–2431. https://doi.org/10.1002/esp.4888

Stafleu, J. & Dubelaar, C.W. (2016) *Product Specification Subsurface Model GeoTOP*. TNO Report R10133. Wageningen: Geological Survey of The Netherlands.

Steur, G.G.L., Heijink, W. & de Bakker, H. (1991) *Bodemkaart van Nederland: schaal 1:50.000/Algemene begrippen en indelingen*, 4e uitg. edition. Wageningen: Staring Centrum.

Stouthamer, E., Cohen, K.M. & Hoek, W. (2020) *De vorming van het land. Geologie en geomorfologie*, 8th edition. Utrecht: Perspectief Uitgevers.

Ten Cate, J.A.M., De Lange, G.W. & Van Maarleveld, G.C. (1975) *Geomorfologische kaart van Nederland: schaal 1: 50.000: Blad 31 Utrecht*. Wageningen: Stichting voor Bodemkartering.

Tobler, W. (1987) Measuring spatial resolution. In: *Proceedings of the Land Resources Information Systems Conference*. Beijing, China, pp. 12–16.

Valentine, A. & Kalnins, L. (2016) An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics. *Earth Surface Dynamics*, 4(2), 445–460. https://doi.org/10.5194/esurf-4-445-2016

Van den Berg, M.W. & Beets, D.J. (1987) Saalian glacial deposits and morphology in The Netherlands. In: Van Der Meer, J.J.M. (Ed.) *Tills and Glaciotectonics*. Rotterdam: Balkema, pp. 235–251.

Van der Meij, W.M. & Maas, G.J. (2020) *Kwaliteitsdocument van de Geomorfologische kaart van Nederland*, Vol. No. 195. Wageningen: Wettelijke Onderzoekstaken Natuur & Milieu. https://doi.org/10.18174/538255

Van der Meulen, B., Cohen, K.M., Pierik, H.J., Zinsmeister, J.J. & Middelkoop, H. (2020) LiDAR-derived high-resolution palaeo-DEM construction workflow and application to the early medieval Lower Rhine valley and upper delta. *Geomorphology*, 370, 107370. https://doi.org/10.1016/j.geomorph.2020.107370

Van der Meulen, M.J., Doornenbal, J.C., Gunnink, J.L., Stafleu, J., Schokker, J., Vernes, R.W. *et al.* (2013) 3D geology in a 2D country: Perspectives for geological surveying in the Netherlands. *Netherlands Journal of Geosciences*, 92(4), 217–241. https://doi.org/10.1017/S0016774600000184

Van Lanen, R.J., Kosian, M.C., Groenewoudt, B.J. & Jansma, E. (2015) Finding a way: Modeling landscape prerequisites for Roman and early-medieval routes in the Netherlands. *Geoarchaeology*, 30(3), 200–222. https://doi.org/10.1002/gea.21510

Verschoof-Van der Vaart, W.B. & Lambers, K. (2019) Learning to look at LiDAR: The use of R-CNN in the automated detection of archaeological objects in LiDAR data from the Netherlands. *Journal of Computer Applications in Archaeology*, 2(1), 31–40. https://doi.org/10.5334/jcaa.32

Verschoof-Van der Vaart, W.B. & Lambers, K. (2021) Applying automated object detection in archaeological practice: A case study from the southern Netherlands. *Archaeological Prospection*, 1–17. https://doi.org/10.1002/arp.1833

Verstappen, H.T. (2011) Old and new trends in geomorphological and landform mapping. In: Smith, M.J., Paron, P. & Griffiths, J.S. (Eds.) *Developments in Earth Surface Processes: Geomorphological Mapping*. Amsterdam: Elsevier, pp. 13–38. https://doi.org/10.1016/B978-0-444-53446-0.00002-1

Wageningen University and Research. (2020) Storymap klimaatopgave Ede. Available at https://geodesk.maps.arcgis.com/apps/MapJournal/index.html?appid=20e4de67e33845fa84e23211431108a4 [accessed 26 October 2021].

Xu, C., Dai, F., Xu, X. & Lee, Y.H. (2012) GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. *Geomorphology*, 145–146, 70–80. https://doi.org/10.1016/j.geomorph.2011.12.040

Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018) Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738

Zagwijn, W.H. (1974) The palaeogeographic evolution of the Netherlands during the Quaternary. *Geologie en Mijnbouw*, 53, 369–385.

Zhang, L., Tanno, R., Xu, M. C., Jin, C., Jacob, J., Ciccarelli, O., Barkhof, F. & Alexander, D. C. (2020) Disentangling Human Error from the Ground Truth in Segmentation of Medical Images. In: *Proceedings of the 34th Conference on Neural Information Processing Systems*. Vancouver, BC, pp. 15750–15762.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F. & Fraundorfer, F. (2017) Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and*

Remote Sensing Magazine, 5(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Wadoux A.M.J.-C., Minasny B., McBratney A.B. (2020) Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: van der Meij, W.M., Meijles, E.W., Marcos, D., Harkema, T.T.L., Candel, J.H.J. & Maas, G.J. (2022) Comparing geomorphological maps made manually and by deep learning. *Earth Surface Processes and Landforms*, 47(4), 1089–1107. Available from: https://doi.org/10.1002/esp.5305